A Rule-Based Method for Outlying Rating Detection

Di Cai and Steve Wade

Abstract—Detection of outlying ratings of samples is a primary step in statistical analysis and classification. A novel rule-based method is presented for automatically detecting and removing outlying ratings in order to improve the quality of sample classification and to increase the degree of agreement between raters. The effectiveness of our method in improving the degree of agreement, assessed using a modified Fleiss' kappa, is demonstrated through a practical example. Our method is conceptually transparent, computationally simple and easy to apply in practice. It is expected to be a useful tool in many real world applications.

Index Terms—Outlying rating detection (ORD), rating frequency distributions, reliability of agreement

I. INTRODUCTION

Sample rating is an important tool and is widely used in industry, psychology, politics and commercial market research, it is also widely used in medical statistics and in food, social and many areas of science. In industry research, many organisations (i.e., banks, telecommunication companies, insurance companies, etc.) track and analyse consumer sentiment for service quality or customer satisfaction [1]. In market research, customers may be asked about their attitudes, perceptions or evaluations of products (or, foods, brands, etc.); managers maybe asked to rate their company's performance (type of strategic focus, degree of marketing excellence, etc.) [1]. Studies [5], [6] show that 20% of data produced by medical researchers in ordered categories; quality assurance in hospitals may result in an increase in the use of methods which produce data in ordered categories [2]. Also, analysis of subjective measurements arises in many research areas, in particular, those concerning sensory testing or attitude scaling [8].

Automatic Outlying Rating Detection (ORD) is an important issue for many real world applications involving sample (data) gathering, analysis, ratings and classification (with ordered categories). A sentiment analysis system and its classifier, for instance, generally rely on the quality of classification of samples in order to accurately predict sentiment orientation of texts or sentences. While many samples, such as reviews rated by web users, are inherently likely to show differences in opinions, a robust and reliable ORD is a prerequisite for effective prediction.

An important area closely related to the current study is outlier detection. The definition of an outlier depends on underlying assumptions regarding the detection method and data structure [2]. Generally, an outlier may be defined as a data point that “appears to deviate markedly from other members of the sample in which it occurs” [1], [6]; or, “lies outside some overall pattern of distribution” [9]. A typical outlier detection technique is to characterise what ‘normal’ data points look like, and then to single out those data points that deviate from these normal properties [15]. There exist many outlier detection methods. A good review of outlier detection methods can be found in, for instance [2], [7].

An outlying rating of a given sample, as referred to in this study, is a rating appearing to deviate significantly from the majority of ratings of the sample. Outlying ratings may arise from experiment design errors and/or human-related errors, unrepresentative assessments or measurements, and so on. Outlying ratings often cause confusion for classification and decrease prediction accuracy. However, the practical and important issue of automatic ORD remains to be developed.

The current study explores a rule-based method for automatic ORD of individual samples. The aim of this pioneering study is to improve the quality of sample classification and to increase the degree of agreement between raters regarding the whole sample set. There are several statistics, for instance, [3], [4], [7], that measure the degree (reliability) of agreement achieved between more than two different raters rating the same samples. Fleiss' kappa measure [3] is simple and commonly used and, thus, it is used in our study. Note that the kappa measure assumes that the number of raters per sample must be fixed, although different samples may be rated by different raters. Therefore, in the current study, we also modify the kappa measure to allow the number of raters to vary from sample to sample. To the best knowledge of the authors, our method is developed for the first time and is expected to be a useful tool for state-of-the-art machine learning methods.

This paper is organized as follows. After giving a notation through a working example in Section II, we introduce a series of basic concepts in Section III. We present a rule-based method for detecting and removing outlying ratings in Section IV and then discuss the extension of our method in Section V. We investigate to what extent our method contributes to increasing the degree of agreement between raters through a practical example in Section VI, and draw conclusions in Section VII.

II. NOTATION

Sample rating, as used in this paper, refers to the process of assigning to each sample some value selected from a list predefined from a given ordered series. The values may be thought of as strength, extent, level, closeness, and so on, depending on the application. In effect, sample rating is equivalent to ‘sample classification’ if each value in the series corresponds to a category. Accordingly, the categories should be clearly defined, ordered and mutually exclusive. In
what follows, we will regard the phrases ‘sample rating’ and ‘sample classification’ as interchangeable.

To begin with, let us give the representation of sample ratings. Suppose we have n samples, denoted by $S = \{s_1, s_2, ..., s_n\}$, of analysis and that we have N values, denoted by $V = \{v_1, v_2, ..., v_N\}$, predefined from a given ordered series. Suppose we have a classification, denoted by $S = \{S_1, S_2, ..., S_k\}$, over $S$, which is a list of ordered categories. We assume that each $s_i \in S$ corresponds to $v_j \in V$, where $j = 1, 2, ..., N$. Suppose there are a total of m raters, and each of them is asked to assign a value to some samples. We call the assignment a rating. In the end, from raters, we obtain arating frequency for each of N values. Then we may represent the ratings of all the samples using an n-by-N table: the samples and values (categories) are arranged in rows and columns, respectively. The table contains $n \times N$ cells, and in the $(i, j)$-th cell contains the rating frequency, denoted by $r_{i,j}$, which is the number of raters who assigned sample $s_i$ with value $v_j$ (or, who classified the $i$th sample $s_i$ into the $j$th category $S_j$). Clearly, table ignores information about raters themselves.

A typical application of ORD is in the area of sentiment analysis [14] and this is where our examples are set. The following working example, Example 2.1 will be used throughout this paper.

**Example 2.1.** Suppose we have a set of $n = 6$ samples, and that there are $m = 30$ raters who were randomly selected and required to classify each sample into $N = 11$ categories. Table I below depicts the statistics (see details in Section VI.B).

<table>
<thead>
<tr>
<th>$S$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$s_2$</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>$s_3$</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$s_4$</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>$s_5$</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>$s_6$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Thus, from Table I, we have $S = \{s_1, s_2, ..., s_6\}$, $V = \{-5, ..., -1, 0, 1, ..., 5\}$ and $S = \{S_1, ..., S_2, S_4, S_7, ..., S_{11}\}$, and the (3, 9)th cell contains the rating frequency, $r_{3,9}$ is 10, which indicates that 10 raters assigned sample $s_3$ with value $v_9 = 3$ (or, classified sample $s_3$ into category $S_9$).

### III. Concepts

This section introduces a series of basic concepts, which are used for characterising the rating frequency distributions of the individual samples. With these concepts, establishing the rules used in our method becomes straightforward.

Generally, there are three variables related to a given sample $s_i \in S$:

1. the number of raters who rated sample $s_i$:

   $$m_i = \sum_{j=1}^{N} r_{i,j} \leq m$$

2. an 1-by-N matrix of rating frequencies of $s_i$ over $S$:

   $$r_i = [r_{i,1}, r_{i,2}, ..., r_{i,j}, ..., r_{i,N-1}, r_{i,N}]$$

3. the distribution of rating frequencies of $s_i$ over $S$:

   $$p_i = \frac{r_i}{m_i} = [p_{i,1}, p_{i,2}, ..., p_{i,j}, ..., p_{i,N-1}, p_{i,N}]$$

For a given $s_i \in S$ with the rating frequency matrix $r_i$ (or, ratings $r_{i,j}$ in short), consider two arbitrary $S_j, S_j' \in S$ (where $j \neq j'$). In current study, the order of the categories in $S$ is necessary. Thus we say $S_j < S_j'$ if $j < j'$. We can define neighbour distance by the following statements.

- The distance between $S_j$ and $S_j'$ is defined by

  $$\text{dis}(S_j, S_j') = |j - j'|$$

- The neighbour distance is a predefined parameter $\alpha$ (a non-negative integer); we say $S_j$ is a neighbour of $S_j'$ if

  $$\text{dis}(S_j, S_j') \leq \alpha$$

A dominant category, which is an important concept of ORD, in this study is the mode of the ratings. That is, a category $S_j \in S$ is said to be a dominant category of sample $s_i \in S$, denoted by $S^*$, if the corresponding rating frequency, denoted by $r_{i,j}$, satisfies

$$r_{i,j} \equiv r_{i,j} = \max(r_{i,j'}, 1 \leq j' \leq N)$$

Obviously, each $s_i$ has at least one dominant category. The main category, which is another important concept of ORD, is the set of neighbouring categories of the dominant category. That is, suppose $S^* = S_j$ is the dominant category of $s_i \in S$, the main category of $S^*$, denoted by $[S^*]$, is defined by

$$[S^*] = \{S_j; \text{dis}(S_j, S^*) \leq \alpha, 1 \leq j \leq N\}$$

where $\alpha$ is neighbour distance.

The left and right main categories of $S^*$ are defined respectively by

$$[S^*]_L = \{S_j; S_j' \in [S^*] \text{ and } 1 \leq j' < j\}$$

$$[S^*]_R = \{S_j; S_j' \in [S^*] \text{ and } j < j' \leq N\}$$

With the above concepts, in what follows, we will denote:

$$\sum_{j \in [S^*]_L} r_{i,j}$$

$$\sum_{j \in [S^*]_R} r_{i,j}$$

which are the sums of rating frequencies over two category sets $[S^*]_L$ and $[S^*]_R$, respectively. They may be viewed as the ‘strengths’ of the left and right neighbours of $S^*$. 

467
IV. OUTLYING RATING DETECTION (ORD)

This section presents our methodology establishing a series of rules for automatic ORD. The basic idea behind our method is simple: the dominant category is the starting point of ORD, which is taken as the highest frequency of ratings for each given sample. There may be alternative ways to define the dominant category, depending on the application. We use the mode for two reasons: (i) the mode captures popularity and, (ii) the mode is insensitive to outlying ratings, except when the number of raters is small (i.e., less than 3). The main category consisting of neighbors of the dominant category, depending on the application. We use ORD, which is taken as the highest frequency of ratings for one dominant category. We will discuss the extension of our method for samples with multiple dominant categories in the next section. That is, for a given \( s \in S \) with the ratings \( r \), suppose \( S^* \) is the unique dominant category of \( s \) over \( S \). The ordered classification \( S \) can thus be expressed by:

\[
S = \{ S_1, ..., S_{i-1}, S^*, S_{i+1}, ..., S_N \}
\]

where \( S^* = S_i \) with \( r_{i,j} = r_{i,j} \).

Let \( \beta \) be a predefined parameter (a non-negative integer), which is the maximum sum of rating frequencies allowed to be removed in our method. Let \( \Gamma_{\text{remove}} \) be the set of rating frequencies currently removed, and denote their sum by

\[
\mu = \sum_{r_{i,j} \in \Gamma_{\text{remove}}} r_{i,j} \geq 0
\]

Let \( \Gamma_{\text{check}} \) be the set of rating frequencies that need to be checked by our method. There are six rules established, denoted by \( R_i \) (\( i = 1, ..., 6 \)), for rating frequency removal:

- For an arbitrary frequency \( r_{i,j} \) \( i \in \Gamma_{\text{check}} \), it is considered to be an outlier and therefore candidate for removal if that rule \( R_1, R_2 \) and \( R_3 \) are simultaneously satisfied:

\[
R_1: \text{dis}(S^*, S^*) > \alpha \\
R_2: r_{i,j} < \frac{1}{2} r_{i,j} \\
R_3: \mu + r_{i,j} < \beta
\]

For an arbitrary frequency \( r_{i,j} \), satisfying \( R_1, R_2 \) and \( R_3 \), it is removed if it further satisfies:

\[
R_4: S_j = \text{arcmax}(\text{dis}(S_j^*, S^*); r_{i,j}^* \in \Gamma_{\text{check}})
\]

- For two arbitrary frequencies \( r_{i,j}^*, r_{i,j}^* \), satisfying:

\[
\begin{align*}
&\text{both } r_{i,j}^* \text{ and } r_{i,j}^* \text{ satisfy } R_4 \quad (a) \\
&S_{i,j}^* < S^* < S_{i,j}^* \quad (b)
\end{align*}
\]

Then, one of the following two rules are applied:

\[
\begin{align*}
R_5: & \quad \text{if } r_{i,j}^* \neq r_{i,j}^* \text{ then} \\
& \quad \text{(a) remove } r_{i,j}^* \text{ if } r_{i,j}^* < r_{i,j}^* \\
& \quad \text{(b) remove } r_{i,j}^* \text{ if } r_{i,j}^* > r_{i,j}^*
\end{align*}
\]

\[
\begin{align*}
R_6: & \quad \text{if } r_{i,j}^* = r_{i,j}^* \text{ then} \\
& \quad \text{(a) remove } r_{i,j}^* \text{ if } (r_{i,j}^*L, L < (r_{i,j}^*)_R) \\
& \quad \text{(b) remove } r_{i,j}^* \text{ if } (r_{i,j}^*_L, L > (r_{i,j}^*)_R) \\
& \quad \text{(c) remove both } r_{i,j}^* \text{ and } r_{i,j}^* \text{ if } (r_{i,j}^*_L, L = (r_{i,j}^*)_R)
\end{align*}
\]

The six rules may be restated:

- \( R_1 \) considers \( r_{i,j} \) as a possible candidate if the correspond \( S_j \) is not a neighbour of \( S^* \);
- \( R_2 \) considers \( r_{i,j} \) as a possible candidate if it is less than half of \( r_{i,j}^* \) of \( S^* \);
- \( R_3 \) considers \( r_{i,j} \) as a possible candidate if it does not result in \( \beta \) being exceeded;
- \( R_4 \) removes \( r_{i,j} \) if \( S_j \) is currently furthest from \( S^* \);
- \( R_5 \) removes the smallest frequency amount \( r_{i,j}^* \) and \( r_{i,j}^* \), if they are not equal to each other (but \( S_j^* \) and \( S_j^* \) have the same furthest distance from \( S^* \));
- \( R_6 \) removes frequency (frequencies) \( r_{i,j}^* \) or and \( r_{i,j}^* \) by means of the strengths of the left and right neighbours of \( S^* \), if they are equal to each other and \( S_j^* \) and \( S_j^* \) have the same furthest distance from \( S^* \).

The six rules should be checked in order and the \( i \)th rule (\( i = 4, 5, 6 \)) above requires all \( R_1 \) to \( R(i-1) \). Let us now see an example below.

Example 4.1. Suppose the neighbour distance \( \alpha = 2 \) and the maximum sum of frequencies allowed to be removed \( \beta = 6 \). Consider sample \( s \) in Table I. For \( S^* = S_6 \), with the corresponding \( S^* = \{ S_4, S_5, S_6, S_7, S_8 \} \), we have

\[
[S^*]_L = \{ S_4, S_5 \} \text{ with } (r_{1,j})_L = r_{5,4} + r_{5,5} = 3 + 6 = 9 \\
[S^*]_R = \{ S_7, S_8 \} \text{ with } (r_{1,j})_R = r_{5,7} + r_{5,8} = 4 + 3 = 7
\]

We initially set \( \Gamma_{\text{removed}} = \emptyset \) and \( \mu = 0 \), where the symbol ‘\( \emptyset \)’ expresses assignment. Then, with rules \( R_1, R_2 \) and \( R_3 \), we have:

- removing \( r_{3,11} \) by \( R_4 \) (\( \mu = r_{5,11} = 0 + 1 = 1 \))
- removing \( r_{3,10} \) by \( R_6(b) \) (\( \mu = r_{5,10} = 1 + 1 = 2 \))
- removing \( r_{3,5} \) by \( R_4 \) (\( \mu = r_{5,5} = 2 + 1 = 3 \))
- removing \( r_{3,5} \) by \( R_4 \) (\( \mu = r_{5,5} = 3 + 1 = 4 \))

We cannot further remove \( r_{5,4} = 3 \) and \( r_{5,8} = 3 \) by \( R_1 \) and \( R_2 \) and, thus, stop after removing \( r_{5,3} = 1 \) with \( \mu = 4 < 6 = \beta \). The removal result for \( s_5 \) is given in Table III (see Section VI.B).

V. EXTENSION

It is likely that samples may have more than one dominant category as several categories may achieve the top rating frequency. For instance, from Table I, we can see that...
the mode of ratings \( r_i \) (of \( s_i \)) is not unique. Our method may be extended to apply to samples with any number of dominant categories.

A. Two Dominant Categories

There are two cases to consider when there are two dominant categories: (i) if the two dominant categories are close to one another, we may generate a ‘new dominant category’ by merging them along with their enclosed neighbours, or (ii) if the two dominant categories are far from each other, we need to split the whole classification (at the midpoint of two modes) into two sub-classifications, each of which contains a single dominant category.

More specifically, for a given \( s_i \in S \) with the ratings \( r_i \), suppose there are two dominant categories \( S^*_1 \) and \( S^*_2 \) over \( S \). The ordered classification \( S \) can be expressed by:

\[
S = \{ S_1, ..., S_{j_1-1}, S^*_1, S_{j_1}, ..., S_{j_2-1}, S^*_2, S_{j_2+1}, ..., S_N \}
\]

where \( S^*_1 = S_{j_1} \) and \( S^*_2 = S_{j_2} \) and \( r_{i,j_1} = r_{i,j_2} \) (\( j_1 \neq j_2 \)). Consider the distance between \( S^*_1 \) and \( S^*_2 \):

\[
dis(S^*_1, S^*_2) = dis(S_{j_1}, S_{j_2}) = |j_2 - j_1|.
\]

Then, we use the neighbour distance to decide whether to split the classification \( S \) as follows:

- If \( dis(S^*_1, S^*_2) \leq 2\alpha \), we say \( S^*_1 \) and \( S^*_2 \) are close to each other. We then generate a new dominant category:

\[
S^* = S^*_1 \cup S_{j_1+1} \cup ... \cup S_{j_2-1} \cup S^*_2
\]

and use the method on

\[
S = \{ S_1, ..., S_{j_1-1}, S^*, S_{j_1+1}, ..., S_N \}
\]

The inequality \( dis(S^*_1, S^*_2) \leq 2\alpha \) is to ensure that the main categories \( [S^*_1] \) and \( [S^*_2] \) are the neighbours, or even have an overlap, of one another.

- If \( dis(S^*_1, S^*_2) > 2\alpha \), we say \( S^*_1 \) and \( S^*_2 \) are far from one another. Let \( \lambda = \lfloor \frac{1}{2} dis(S^*_1, S^*_2) \rfloor \) and take the floor function \( \lambda = \lfloor x \rfloor \) (\( x \) is the floor function, which is the largest integer not greater than \( x \)) and, then, use our method twice, on

\[
S = \{ S_1, ..., S_{j_1-1}, S^*_1, S_{j_1+1}, ..., S_{j_1+\lambda} \}
\]

\[
S_2 = \{ S_{j_1+\lambda+1}, ..., S_{j_2-1}, S^*_2, S_{j_2+1}, ..., S_N \}
\]

Clearly, \( 0 \leq \lambda \leq \frac{N}{2} \), that is, \( \lambda \) reaches the maximum if \( S^*_1 = S_{j_1} \) and \( S^*_2 = S_{j_2} \).

Note that the sum of the ratings removed from two sub-classifications \( S_1 \) and \( S_2 \) should not exceed \( \beta \).

B. More Than Two Dominant Categories

For a given \( s_i \in S \) with ratings \( r_i \), suppose there are more than two dominant categories over \( S \). Let us denote \( \Omega_i \) as the set of all the dominant categories of \( s_i \):

\[
\Omega_i = \{ S^*_1, ..., S^*_i, ..., S^*_T \}
\]

where \( \tau_i = |\Omega_i| \) is the size of \( \Omega_i \). That is, \( s_i \) has \( \tau_i \) dominant categories, \( S^*_i \) = \( S_l \) with ratings \( r_{ij} \) (\( l = 1, 2, ..., \tau_i \)) over \( S \). Then the ordered classification \( S \) can be expressed by:

\[
S = \{ S_1, ..., S_{j_1-1}, S^*_1, S_{j_1}, ..., S_{j_2-1}, S^*_2, S_{j_2-1}+1, ..., S_N \}
\]

For each dominant category pair \((S^*_l, S^*_l+1)\), consider the distance \( dis(S^*_l, S^*_l+1) \) successively, where \( l = 1, 2, ..., \tau_i - 1 \), and apply our method for the case where there are only two dominant categories.

VI. Effectiveness

This section concentrates on the effectiveness of our method. As mentioned previously, the aim of this study is to improve the quality of sample classification and to increase the degree of agreement between raters. Therefore, we investigate to what extent our method contributes to the increase through a practical example. We first modify the statistical measure Fleiss’ kappa and, then calculate the degree obtained from our method and compare them with the original degree without outlying rating removal.

A. Fleiss’ Kappa

The degree of agreement obtained from the original Fleiss’ kappa [3], denoted by \( \kappa \), is a real number. It assumes that the number of the raters per sample is fixed when assigning category ratings to a number of samples. Note that, after applying our method, it is very likely that data is incomplete as some cells in the resultant table are ‘empty’ (see Table III below). Thus the number \( m \) may vary from sample to sample. Therefore, the estimate of probabilities (or, proportions \( p_{and,p_j} \) required in the measure \( \kappa \) should be modified to allow the number of raters to vary from sample to sample. We modify the estimate and denote the modified measure by \( \bar{\kappa} \).

On one hand, the degree of agreement among the \( m \) raters for sample \( s_i \) can be expressed by the proportion of agreeing pairs out of all the \( m(m-1)/2 \) possible pairs of assignments:

\[
P_i = \frac{1}{m_i(m_i-1)/2} \sum_{j=1}^{m_i} r_{ij}(r_{ij} - 1)
\]

\[
= \frac{1}{m_i(m_i-1)} [\sum_{j=1}^{m_i} r_{ij}^2 - m_i]
\]

(1)

where \( m_i(m_i-1) \) may be viewed as a normalization factor. The average degree of agreement thus expressed by:

\[
\bar{P} = \frac{1}{n} \sum_{i=1}^{n} P_i
\]

(2)

which means, if \( s_i \) was classified by two randomly selected raters, that the (average) probability of the second rater agreeing with the first is \( \bar{P} \).

On the other hand, the proportion of all assignments, for instance, to category \( S_j \) is:

\[
p_{ij} = \frac{1}{M'} \sum_{i=1}^{n} r_{ij} = \frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} r_{ij}
\]

(3)

where \( M' \) is a normalization factor, which is the sum of
ratings in the individual cells. Thus, if all the raters made their assignments purely by chance, the mean proportion of agreement over the classification should be:

$$\bar{P}_e = \sum_{j=1}^{n} p_{e,j}^2$$  \hspace{1cm} (4)

Finally, we have the modified measure:

$$\kappa' = \frac{P - P_e}{1 - P_e}$$  \hspace{1cm} (5)

where $P - P_e$ is the degree of agreement actually attained from ratings in excess of chance; $1 - P_e$ is the degree attainable above what would be predicted by chance.

It is worth mentioning, when $m_i = m$ ($i = 1, 2, ..., n$), that we have $\kappa' = \kappa$. That is, $\kappa$ is a special case of $\kappa'$.

B. Application Example

In the area of sentiment analysis [14], users are instructed to rate comments (on some produce), extracted from web blogs, for strength of negative/positive sentiment. The sentiment strengths (SS) on an 11-point scale are given in Table II: the points -5 to -1 are from very strong negative to weak negative; the points 1 to 5 are from weak positive to very strong positive; the point 0 is both neutral and ‘do not know’ (or, ‘undecided’).

Six matrices of ratings for 6 comments (i.e., samples $s_1$ to $s_6$), obtained from $m = 30$ users (i.e., raters), are shown in Table I (see Example 2.1 in Section II). Table III below shows the results, after applying our method to the 6 samples, in which, ratings with * * indicate the corresponding category is the dominant category, and a hyphen indicates removed ratings.

### Table III: After Applying Ord

<table>
<thead>
<tr>
<th>S</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
<th>-1</th>
<th>-2</th>
<th>-3</th>
<th>-4</th>
<th>-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td><em>9</em></td>
<td>5</td>
<td>24</td>
</tr>
<tr>
<td>$s_2$</td>
<td>1</td>
<td>7</td>
<td><em>8</em></td>
<td>4</td>
<td>2</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
</tr>
<tr>
<td>$s_3$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
</tr>
<tr>
<td>$s_4$</td>
<td>3</td>
<td><em>7</em></td>
<td>4</td>
<td><em>7</em></td>
<td>3</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
</tr>
<tr>
<td>$s_5$</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>6</td>
<td><em>10</em></td>
<td>4</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
</tr>
<tr>
<td>$s_6$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td><em>9</em></td>
<td>-</td>
<td>-</td>
<td>27</td>
</tr>
</tbody>
</table>

Total($S_i$) | 4 | 14 | 12 | 14 | 11 | 12 | 10 | 12 | 22 | 22 | 18 |

$p_{i,j}$ | 0.0265 | 0.0927 | 0.0795 | 0.0927 | 0.0728 | 0.0795 | 0.0662 | 0.0795 | 0.1457 | 0.1457 | 0.1192 |

Note that, from Table III, we have $M' = \sum_{i=1}^{n} m_i = 151$. Thus, with (1) and (3), taking the first row and last column, for instance, we have

$$P_1 = \frac{1}{24(24 - 1)}(3^2 + 1^2 + 2^2 + \cdots + 5^2 - 24) = 0.2319$$

$$P_{11} = \frac{1}{151}(5 + 4 + 9) = 0.1192$$

Then, with (2) and (4), we have

$$\bar{P} = \frac{1}{6}(0.2319 + 0.2065 + \cdots + 0.2336) = 0.2188$$

$$\bar{P}_e = 0.0265^2 + 0.0927^2 + \cdots + 0.1192^2 = 0.1032$$

Finally, with (5), we obtain

$$\kappa' = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} = \frac{0.2188 - 0.1032}{1 - 0.1032} = 0.1289$$

For comparison, for the 6 samples given in Table I, with the original Fleiss’ kappa [3], the degree of agreement among them = 30 raters for sample $s_1$, for instance, can be expressed by

$$P_1 = \frac{1}{m(m - 1)}\left(\sum_{j=1}^{n} r_{i,j}^2 - m\right)$$

$$= \frac{1}{30(30 - 1)}[(12 + 0^2 + \cdots + 9^2 + 5^2 - 30) = 0.1517$$

### Table IV: Before Applying Ord

<table>
<thead>
<tr>
<th>S</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
<th>-1</th>
<th>-2</th>
<th>-3</th>
<th>-4</th>
<th>-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>9</td>
<td>5</td>
<td>30</td>
<td>0.1517</td>
</tr>
<tr>
<td>$s_2$</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>30</td>
<td>0.1333</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>30</td>
<td>0.1632</td>
</tr>
<tr>
<td>$s_4$</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>30</td>
<td>0.1287</td>
</tr>
<tr>
<td>$s_5$</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>30</td>
<td>0.1655</td>
</tr>
<tr>
<td>$s_6$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>30</td>
<td>0.1885</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total($S_i$) | 5 | 16 | 17 | 16 | 14 | 12 | 15 | 24 | 25 | 20 | 180 |

$p_{i,j}$ | 0.0278 | 0.0889 | 0.0944 | 0.0889 | 0.0889 | 0.0778 | 0.0667 | 0.0833 | 0.1333 | 0.1389 | 0.1111 |
Thus, the proportion of all assignments, for instance, to category $S_{11}$ is:

$$p_{r,11} = \frac{1}{M} \sum_{i=1}^{n} r_{i,11} = \frac{1}{180} (5 + 1 + \cdots + 1 + 9) = 0.1111$$

where $M = m \times n = 180$. Then, we obtain

$$\bar{p} = \frac{1}{6} (0.1517 + 0.1333 + \cdots + 0.1885) = 0.1552$$

$$\bar{p_e} = 0.0278^2 + 0.0889^2 + \cdots + 0.1111^2 = 0.1002$$

Finally, we obtain

$$\kappa = \frac{0.1552 - 0.1002}{1 - 0.1002} = 0.0610$$

Thus, from the above results, we can see $\kappa' - \kappa = 0.1289 - 0.0610 = 0.0679$. That is, we obtain a 6.79% increase in the degree of agreement after applying our method. Note that the above $m(m - 1)$ and $M$ are the normalization factors, which are different from ones given in (1) and (3), respectively.

VII. CONCLUSION

This study explored a novel method for automatically detecting and removing outlying ratings. A series of basic concepts were introduced, which are used to characterise the rating frequency distributions and to establish rules for detecting and removing outliers. The key point of our method is that the rating frequency is regarded as an outlier and removed if (i) it exhibits a very low frequency and/or, (ii) a high divergence from the mode. Our method was presented for samples with a single dominant category; it was also extended to samples with multiple dominant categories. The effectiveness of our method in improving the degree of agreement between raters, assessed with the modified Fleiss' kappa, was demonstrated through a practical example. It should be pointed out that the rating frequency distributions of samples may be very complex and that the current study is pioneering work, so there remains a large gap to be filled for future work. Finally, we would expect ORD to be a useful tool in real-world applications, in particular, involving web data gathering, ratings and classification.

REFERENCES