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 marketingexcellence, etc.) [1]. Studies [5], [6] show that 20% of data produced by medical researchis in ordered categories; quality assurance in hospitals may result in an increase in the useof methods which produce data in ordered categories [2]. Also, analysis of subjective measurements arises in many research areas, in particular, those concerning sensory testing orattitude 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 assumptionsregarding 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 todeviate significantly from the majority of ratings of the sample. Outlying ratings may arisefrom experiment design errors and/or human-related errors, unrepresentative assessmentsor measurements, and so on. Outlying ratings often cause confusion for classification anddecrease prediction accuracy. However, the practical and important issue of automatic ORDremains 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 toincrease the degree of agreement between raters regarding the whole sample set. There are severalstatistics, for instance, [3], [4], [7], that measure the degree (reliability) of agreement achievedbetween 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 differentsamples may be rated by different raters. Therefore, in the current study, we also modify the *kappa* measure toallow the number of raters to vary from sample to sample. To the best knowledge of theauthors, our method is developed for the first time and is expected to be a useful toolfor state-of-the-art machine learning methods.

This paper is organized as follows. After giving a notation through aworking 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 toincreasing 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 somevalue selected from a list predefined from a given ordered series. The values may be thoughtof 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 correspondsto a category. Accordingly, the categories should be clearly defined, ordered and mutuallyexclusive. In

Manuscript received July 30, 2012; revised September 2, 2012.

The authors are with School of Computing and Engineering, University of Huddersfield, HD1 3DH, UK (e-mail: d.cai@hud.ac.uk, s.j.wade@hud.ac.uk)

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 *nsamples,* denoted by $S = \{s_1, s_2, \dots, s_n\}$, of analysis and that we have *Nvalues*, denoted by $V = \{v_1, v_2, \dots, v_N\}$, predefined from a given orderedseries. Suppose we have a classification, denoted by $\boldsymbol{S} = \{S_1, S_2, \dots, S_N\}$, over S, which is a list of ordered categories. We assume that each$S_j \in \boldsymbol{S}$correspondsto$v_j \in V$, where$j = 1,2, \dots, N$. Suppose there are a total of *mraters*, and each of them is asked to assign a valueto some samples. We call the assignment a *rating*. In the end, from *m*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) arepresented in rows and columns, respectively. The table contains $n \times N$cells, and the (i, j)thcell 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 areset. The following working example, Example 2.1 will be usedthroughout 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 I: RATING FREQUENCIES FOR AN ORDERED CLASSIFICATION

	-5	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**
$S \setminus \boldsymbol{S}$	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}
s_1	1		2	1	2		3	1	6	9	5
s_2	1	7	8	4	2	1	2	1	2	1	1
s_3		1	2	1	2		1	3	10	6	4
s_4	3	7	4	7	3	2	1	2		1	
s_5		1	1	3	6	10	4	3		1	1
s_6					1	1	1	5	6	7	9

Thus, from Table I, we have$S = \{s_1, s_2, \dots, s_6\}$, $V = \{-5, \dots, -1, 0, 1, \dots, 5\}and\boldsymbol{S} = \{S_1, \dots, S_5, S_6, S_7, \dots, S_{11}\}$, and the (3, 9)th cell contains the rating frequency, $r_{3,9}$ = 10, which indicates that 10 ratersassigned 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 ratingfrequency 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) an1-by-N matrix of rating frequencies of s_i over $\boldsymbol{S}$:

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

3) the distribution of rating frequencies of s_i over $\boldsymbol{S}$:

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

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

- The *distance* betweenS_j and $S_{j'}$ is defined by

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

- The *neighbour distance* is a predefined parameter α(a non-negative integer); we say $S_{j'}$is a neighbour of $S_j$if

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

A dominant category, which is an important concept of ORD, in this study is the *mode* of the ratings r_i.That is, a category$S_j \in \boldsymbol{S}$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

$$r^*_{i,j} \triangleq r_{i,j} = max\{r_{i,j'}; 1 \leq j' \leq N\}$$

Obviously, each s_i has at least one dominant category.

The main category, which is another important concept of ORD, is the set of neighbouringcategories 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

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

where α is neighbour distance.

The *left* and *right* main categories of S^* are definedrespectively 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:

$$(r^*_{i,j})_L = \sum_{S_{j'} \in [S^*]_L} r_{i,j'}$$

$$(r^*_{i,j})_R = \sum_{S_{j'} \in [S^*]_R} r_{i,j'}$$

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^*.

IV. OUTLYING RATING DETECTION (ORD)

This section presents our methodby establishing a series of rules for automatic ORD. The basic idea behind our method is simple: the dominant category is astarting point of ORD, which is taken as the highest frequency of ratings for eachgiven sample. There may be alternative ways to define the dominant category, depending onthe 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 neighbours of thedominant category is regarded as the range of ratings considered 'normal'. Then, 'abnormal' ratings are regarded as outliers. The outliersare removed, based on the rules established,if they exhibit very low frequencies and/or a high divergence from the dominant category.

At the moment, we consider those samples having only 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$s_i \in S$ with the ratings r_i, suppose S^* is the unique dominant category of s_i over $\boldsymbol{S}$. The orderedclassification $\boldsymbol{S}$ can thus be expressed by:

$$\boldsymbol{S} = \{S_1, \dots, S_{j-1}, S^*, S_{j+1}, \dots, S_N\}$$

where $S^* = S_j$ with $r^*_{i,j} = r_{i,j}$.

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

$$\mu = \sum_{r_{i,j'} \in \Gamma_{removed}} r_{i,j'} \geq 0$$

Let Γ_{check}be the set of rating frequencies that needto be checked by our method. There are six rules established, denotedby Ri ($i = 1, \dots, 6$), for rating frequency removal:

- For each frequency$r_{i,j'} \in \Gamma_{check}$, it is considered to be an outlier and thereforea candidate for removal if three rules R1, R2 and R3 are simultaneously satisfied:

$$\text{R1: } dis(S_{j'}, S^*) > \alpha$$

$$\text{R2: } r_{i,j'} < \frac{1}{2} r^*_{i,j}$$

$$\text{R3: } \mu + r_{i,j'} \leq \beta$$

For an arbitrary frequency$r_{i,j'}$, satisfying R1, R2 and R3, it is removed if it further satisfies:

$$\text{R4: } S_{j'} = arcmax\{dis(S_{j''}, S^*); r_{i,j''} \in \Gamma_{check}\}$$

- For two arbitrary frequencies r_{i,j'_1}, r_{i,j'_2}, satisfying:

  (a) both r_{i,j'_1} and r_{i,j'_2} satisfy R4

  (b) $S_{j'_1} < S^* < S_{j'_2}$

Then, one of the following two rules are applied:

R5: if $r_{i,j'_1} \neq r_{i,j'_2}$ then
  (a) remove r_{i,j'_1} if $r_{i,j'_1} < r_{i,j'_2}$
  (b) remover_{i,j'_2} if $r_{i,j'_1} > r_{i,j'_2}$

R6: if $r_{i,j'_1} = r_{i,j'_2}$ then
  (a) remove r_{i,j'_1} if $(r^*_{i,j})_L < (r^*_{i,j})_R$
  (b) remove r_{i,j'_2} if $(r^*_{i,j})_L > (r^*_{i,j})_R$
  (c) remove both r_{i,j'_1} and r_{i,j'_2}if $(r^*_{i,j})_L = (r^*_{i,j})_R$

The six rules may be restated:

- R1 considers $r_{i,j'}$ as a possiblecandidate if the correspond $S_{j'}$ is not a neighbour of S^*;
- R2 considers $r_{i,j'}$as a possiblecandidate if it is less than half of $r^*_{i,j}$ of S^*;
- R3 considers $r_{i,j'}$as a possiblecandidate ifits removal does not result in β being exceeded;
- R4 removes $r_{i,j'}$ if $S_{j'}$ is currently furthest from S^*;
- R5 removes the smallest frequency amount r_{i,j'_1} and r_{i,j'_2}, if they are not equal to each other (but $S_{j'_1}$and $S_{j'_2}$have the same furthest distance from S^*);
- R6removes frequency (frequencies) r_{i,j'_1} or/and r_{i,j'_2} by means of the strengths of the left and right neighbours of S^*, if they are equal to each other (and$S_{j'_1}$ and $S_{j'_2}$have thesame furthest distance fromS^*).

The six rules should be checked in order and the ith rule (i = 4,5,6) above requires all R1 toR(i-1). Let us now see an example below.

Example 4.1. Suppose the neighbour distance $\alpha = 2$ and themaximum sum of frequencies allowed to be removed $\beta = 6$. Consider sample s_5 in Table I. For $S^* = S_6$with the corresponding$[S^*] = \{S_4, S_5, S_6, S_7, S_8\}$, we have

$$[S^*]_L = \{S_4, S_5\} \text{with} (r^*_{5,6})_L = r_{5,4} + r_{5.5} = 3 + 6 = 9$$

$$[S^*]_R = \{S_7, S_8\} \text{with} (r^*_{5,6})_R = r_{5,7} + r_{5.8} = 4 + 3 = 7$$

We initially set $\Gamma_{removed} \Leftarrow \Phi$ and $\mu \Leftarrow 0$, where the symbol '$\Leftarrow$' expresses assignment. Then, with rulesR1, R2 and R3, we have:

- removing$r_{5,11}$by R4 ($\mu \Leftarrow \mu + r_{5,11} = 0 + 1 = 1$)
- removing $r_{5,10}$by R6(b) ($\mu \Leftarrow \mu + r_{5,10} = 1 + 1 = 2$
- removing $r_{5,2}$by R4 ($\mu \Leftarrow \mu + r_{5,2} = 2 + 1 = 3$
- removing $r_{5,3}$by R4 ($\mu \Leftarrow \mu + r_{5,3} = 3 + 1 = 4$)

We cannot further remove $r_{5,4} = 3$ and $r_{5,8} = 3$ by R1 and R2 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 mayachieve the top rating frequency. For instance, from Table I, we can see that the mode of ratings r_4 (of s_4) is not unique. Our method may beextended 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 dominantcategories are close to one another, we may generate a 'new dominant category' by merging themalong 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 $\boldsymbol{S}$. The ordered classification $\boldsymbol{S}$ can be expressed by:

$$\boldsymbol{S} = \{S_1, \dots, S_{j_1-1}, S_1^*, S_{j_1+1}, \dots, S_{j_2-1}, S_2^*, S_{j_2+1}, \dots, S_N\}$$

where $S_1^* = S_{j_1}$, $S_2^* = S_{j_2}$ and $r_{i,j_1} = r_{i,j_2}$ ($j_1 \neq j_2$). Consider the distance betweenS_1^* andS_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 $\boldsymbol{S}$ 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:

  $$S^* = S_1^* \cup S_{j_1+1} \cup \dots \cup S_{j_2-1} \cup S_2^*$$

  and use the method on

  $$\boldsymbol{S} = \{S_1, \dots, S_{j_1-1}, S^*, S_{j_2+1}, \dots, S_N\}$$

  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 $\lambda' = \frac{1}{2} dis(S_1^*, S_2^*)$ and take the floor function $\lambda = \lfloor \lambda' \rfloor$ ($\lfloor x \rfloor$is the floor function, which is the largest integer not greater than x) and, then, use our method twice, on

  $$\boldsymbol{S}_1 = \{S_1, \dots, S_{j_1-1}, S_1^*, S_{j_1+1}, \dots, S_{j_1+\lambda}\}$$

  $$\boldsymbol{S}_2 = \{S_{j_1+(\lambda+1)}, \dots, S_{j_2-1}, S_2^*, S_{j_2+1}, \dots, S_N\}$$

Clearly, $0 \leq \lambda' \leq \frac{N}{2}$, that is, λ' reaches themaximum if$S_1^* = S_1$ and $S_2^* = S_N$.

Note that the sum of the ratings removed from two sub-classifications$\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ should not exceed β.

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$\boldsymbol{S}$. Let us denote Ω_i as the set of all the dominant categories of s_i:

$$\Omega_i = \{S_1^*, \dots, S_l^*, \dots, S_{\tau_i}^*\}$$

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

$$\boldsymbol{S} = \{S_1, \dots, S_{j_1-1}, S_1^*, S_{j_1+1}, \dots, S_{j_l-1}, S_l^*, S_{j_l+1}, \dots, S_{j_{\tau_i}-1}, S_{\tau_i}^*, S_{j_{\tau_i}+1}, \dots, S_N\}$$

For each dominant category pair (S_l^*, S_{l+1}^*),consider the distance $dis(S_l^*, S_{l+1}^*)$ successively, where $l = 1,2, \dots, \tau_i - 1$, 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, theaim of this study is to improve the quality of sample classification and to increase the degreeof agreement between raters. Therefore, we investigate to what extent our method contributes tothe increase through a practical example. We first modify statistical measureFleiss' *kappa* and, then calculate the degree obtained from our method and compare them withthe original degree without outlying rating removal.

A. Fleiss' Kappa

The degree of agreement obtained from the original Fleiss' *kappa* [3], denoted by κ, is a realnumber. It assumes that the number of the raters per sample is fixed when assigning categoryratings to a number of samples. Note that, after applying our method, it is very likely thatdata is incomplete as some cells in the resultant table are 'empty' (see Table III below). Thus the number m_i may vary from sample to sample.Therefore, the estimate of probabilities (or, proportions $P_i$and$p_{\cdot,j}$) required in the measure κ should be modified to allowthe number of raters to vary from sample to sample. We modify the estimate and denote the modifiedmeasure byκ'.

On one hand, the degree of agreement among the $m_i$raters for sample s_i may be expressedby the proportion of agreeing pairs out of all the $m_i(m_i - 1)$possible pairs of assignments:

$$P_i = \frac{1}{m_i(m_i-1)} \sum_{j=1}^{N} r_{i,j}(r_{i,j} - 1) = \frac{1}{m_i(m_i-1)} \left[\left(\sum_{j=1}^{N} r_{i,j}^2 \right) - m_i \right] \quad (1)$$

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

$$\bar{P} = \frac{1}{n} \sum_{i=1}^{n} P_i \quad (2)$$

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:

$$p_{\cdot,j} = \frac{1}{M'} \sum_{i=1}^{n} r_{i,j} = \frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} r_{i,j} \quad (3)$$

whereM' 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 proportionof agreement over the classification should be:

$$\bar{P}_e = \sum_{j=1}^{n} p_{\cdot,j}^2 \qquad (4)$$

Finally, we have the modified measure:

$$\kappa' = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e} \qquad (5)$$

where $\bar{P} - \bar{P}_e$ is the degree of agreement actually attained from ratings in excess ofchance; $1 - \bar{P}_e$ is the degree attainable above what would be predicted bychance.
It is worth mentioning, when $m_i = m$ $(i = 1,2, \ldots, n)$, that we have $\kappa' = \kappa$. That is, κ is a special case of κ'.

B. Application Example

In the area of sentiment analysis [14], users are instructed to ratecomments (on some produce), extracted from web blogs, for strength of negative/positivesentiment. 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 weakpositive 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 IIIbelow 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 II: SENTIMENT STRENGTHS ON AN 11-POINT SCALE

Value (SS)	Description
-5	very strong negative sentiment
-4	strong negative sentiment
-3	not very strong negative sentiment
-2	mild negative sentiment
-1	weak negative sentiment
0	neutral
1	weak positive sentiment
2	mild positive sentiment
3	not very strong positive sentiment
4	strong positive sentiment
5	very strong positive sentiment

TABLE III: AGREEMENT AFTER APPLYING ORD

	-5	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**		
$S \setminus \mathbf{S}$	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	m_i	P_i
s_1	-		-	-	-		3	1	6	*9*	5	24	0.2319
s_2	1	7	*8*	4	2	-	2	-	-	-	-	24	0.2065
s_3		-	-	-	-		1	3	*10*	6	4	24	0.2500
s_4	3	*7*	4	*7*	3	2	-	-		-		26	0.1692
s_5		-	-	3	6	*10*	4	3		-	-	26	0.2215
s_6					-	-	-	5	6	7	*9*	27	0.2336
Total(S_j)	4	14	12	14	11	12	10	12	22	22	18	151	
$p_{\cdot,j}$	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

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

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

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 givenin Table I, with the original Fleiss' *kappa* [3], the degree of agreement among the$m = 30$raters forsample s_1, for instance, can be expressed by

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

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

TABLE IV: AGREEMENT BEFORE APPLYING ORD

	-5	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**		
$S \setminus \mathbf{S}$	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	m_i	P_i
s_1	1		2	1	2		3	1	6	9	5	30	0.1517
s_2	1	7	8	4	2	1	2	1	2	1	1	30	0.1333
s_3		1	2	1	2		1	3	10	6	4	30	0.1632
s_4	3	7	4	7	3	2	1	2		1		30	0.1287
s_5		1	1	3	6	10	4	3		1	1	30	0.1655
s_6					1	1	1	5	6	7	9	30	0.1885
Total(S_j)	5	16	17	16	16	14	12	15	24	25	20	180	
$p_{\cdot,j}$	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:

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

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 aredifferent 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 ratingfrequency distributions and to establish rules for detecting and removing outliers. The key point of our method is that arating 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 alsoextended to samples with multiple dominant categories. The effectiveness of our method in improving thedegree of agreement between raters, assessed with the modified Fleiss' *kappa*, wasdemonstrated through a practical example. It should be pointed out that the rating frequencydistributions 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 bea useful tool in real world applications, in particular, involving web data gathering, ratingsand classification.

References

[1] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed.: Wiley-Blackwell, 1994.

[2] I. B. Gal, "Outlier Detection," in *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, O. Maimon and L. Rockach: Kluwer Academic Publishers, 2005, pp. 131-146.

[3] J. Dawes, "Do data characteristics change according to the number of scale points used?" *International Journal of Market Research*, vol. 50, no. 1, pp. 61-77, 2008.

[4] A. Donabedian, "The quality of care; how can it be assessed," *JAMA*, vol. 260, no. 12, pp. 1743-1748, Sep. 1988.

[5] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378-382, Nov. 1971.

[6] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1-12, Nov. 1969.

[7] T. Hu and S. Y. Sung, "Detecting pattern-based outliers," *Pattern Recognition Letters*, vol. 24, no. 16, pp. 3059-3068, Dec. 2003.

[8] K. Krippendorff, *Content Analysis: An Introduction to its Methodology*, Thousand Oaks, CA: Sage Publications, 2004.

[9] D. S. Moore and G. P. McCabe, *Introduction to the Practice of Statistics*: W. H. Freeman and Company, 1999.

[10] A. P. Morton and A.J. Dobson, "Analysing ordered categorical data from two independent samples," *British Medical Journal (BMJ)*, vol. 301, no. 6758, pp. 971-973, Oct. 1990.

[11] L. Moses, J. Emerson, and H. Hosseini, "Analysing data from ordered categories," *The New England Journal of Medicine*, vol. 311, no. 7, pp. 442-448, 1984.

[12] J. Sim and C.C. Wright, "The kappa statistic in reliability studies: Use, interpretation, and sample size requirements," *Physical Therapy*, vol. 85, no. 3, pp. 257-268, Mar. 2005.

[13] E. J. Snell, "A scaling procedure for ordered categorical data," *Biometrics*, vol. 20, no. 3, pp. 592-607, Sep. 1964.

[14] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544-2558, Dec. 2010.

[15] J. A. Ting, A. D'Souza, and S. Schaal, "Automatic outlier detection: A Bayesian approach," in *IEEE International Conf. Robotics and Automation*, 2007, pp. 2489-2494.

Di Cai received her PhD in the Department of Computing Science at the University of Glasgow in UK. She is currently a research fellow in the School of Computing and Engineering at the University of Huddersfield in UK. Her main research interests include information extraction and retrieval, document classification and summarization, text mining and analytics, emotion and sentiment analysis. She is a member of the IEEE and ACM.

Steve Wade is a senior lecturer in Information Systems in the Department of Informatics at the University of Huddersfield in UK, where he has been since 1993. He teaches Web Programming and Database Development and has research interests in information retrieval and information systems development methods.