Journal of Information Science
37(3) 235–245
© The Author(s) 2011
Reprints and permission: sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0165551511400954
jis.sagepub.com
SAGE

An automatic recognition method of journal impact factor manipulation

Guang Yu, Dong-Hui Yang
School of Management, Harbin Institute of Technology, P. R. China

Hui-Xin He
School of Energy, Harbin Institute of Technology, P. R. China

Abstract

Journal impact factor (IF) manipulation has unhealthy effects on the academic community and is attracting more attention from scholars. In this paper, an intelligent method is proposed to identify manipulative self-citation behaviour in journals using pattern recognition. Data on IFs, age distributions of total citations, and numbers of self-citations were collected for 18 journals from 1998 to 2007 in Journal Citation Reports (JCR); these journals include known manipulated journals. The feature variables of the citation distribution functions of the known manipulated journals were extracted using the *k*-nearest neighbour classifier, and a feature attribute space was established for pattern recognition. The MATLAB software was used to process, train, and test the data and to develop a suitable matrix model which can provide an original model for identifying other manipulated journals. To verify the validity and reliability of this method, the authors randomly collected citation distribution data from several journals in JCR, analysed the results of the verification, and proved the effectiveness of pattern recognition in this context.

Keywords

impact factor; manipulation; abnormal self-citation; pattern recognition; *k*-nearest neighbour

1. Introduction

Impact factor (IF) is a measure of the influence of a journal and is an important measurement for evaluating the individual careers of researchers and scientists. It is almost exclusively used worldwide by universities, research institutions, and governmental agencies for judging the impact of a professional journal [1]. For this reason, every editor tries to increase the IF of their journal. Because of limitations in its calculation procedure, the IF can easily be manipulated [2]. References to some journals can be controlled, which leads to an increase in the number of the journals cited in the previous two years and consequently an increase in the IF [3]. Thus, these journals will be seen to have additional impact, and the quantity and quality of their journal contributors will be increased in proportion. Unfortunately, the real influence of these journals is not reinforced. To some extent, this manipulative behaviour will lead to artificial evaluation results.

If self-citations of a journal were artificially increased over a long period, the age distribution of self-citations would inevitably disagree with the citation age distribution of this journal. Manipulative behaviour will bring about abnormal characteristics in the self-citation model. Fowler and Aksnes found that the greater the number of self-citations in a volume, the greater the total number of citations for this volume [4]. Moreover, within the next few years, an increasing citation volume for the journal would be observed. As a presentation format for the development of science, which has its own rules, a journal also has its own rules. However, manipulation breaks the rules, which is bound to have a negative impact on journal development.

Corresponding author:
Guang Yu, Library, Harbin Institute of Technology, No. 17, Silin Street, NanGang, Harbin, 150001, P.R. China.
Email: yug@hit.edu.cn

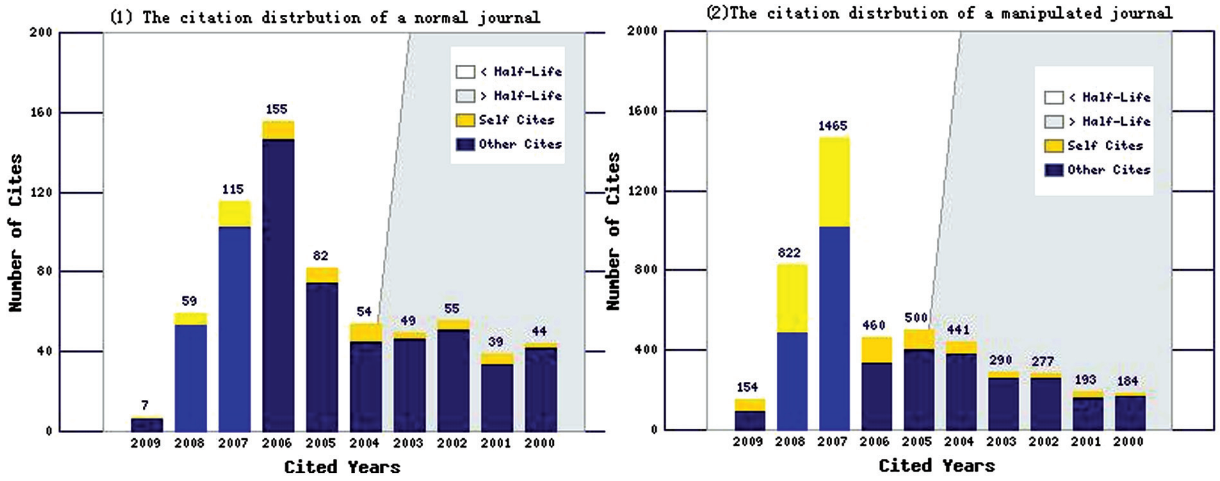

Figure 1. Citation distribution charts of two types of journals from the *JCR* database

In recent years, there have been many studies on self-citation worldwide, addressing topics such as the rule parameters for journal self-citation, comparisons of journal self-citation rules in different subject areas, and the effects on the IF of journal self-citation [5,6]. Other researchers have focused on IF manipulation [3,7–9]. These researchers have analysed the reasons for manipulation and condemned this behaviour. Falagas and Alexiou (2008) described the top 10 journals involved in journal IF manipulation [10]. Krell suggested that authors are becoming increasingly exercised if there is any sign that IFs are being manipulated and that editors who ask authors to cite relevant papers from their own journal are now being accused of acting unethically [11]. Yu and Wang developed a mathematical expression of the relation between a journal's self-citation rate and its IF and analysed the possibility that journal editors manipulate the IFs of their journals by raising the self-citation rate [8]. Yu et al. studied various effects between the IF and the reliability-based citation IF (shortened form: R-IF) in the manipulation process and found that the R-IF is fairer than the IF for journals with relatively short citation half-lives [11]. According to journal IF manipulation research, an increase in IF caused by manipulative self-citation is usually accompanied by changes in the parameters of the journal citation distribution. In other words, the journal citation process of the journal is biased by factors such as IF growth rate and citation half-life [12]. The citation pattern of each manipulated journal would be abnormal, especially the number of self-citations in the previous two years. Figure 1 shows citation distribution charts of two types of journal. We can see that the self-cited rate of the normal journal is lower than that of the manipulated journal. Therefore, it is possible to identify abnormal journals through changes in those parameters. However, it is a monumental task to characterize manually each journal in the *Journal Citation Report (JCR)* database. This paper therefore proposes a pattern recognition approach using machine intelligence to identify suspect journals.

In recent years, pattern recognition has been used for a wide range of information processing problems in many fields. These problems include speech recognition, handwritten character recognition, face recognition, medical research, machine vision, target recognition, radar image analysis, geographical information processing, financial time series prediction and text analysis. However, no one has yet applied pattern recognition to journal classification and identification. In this paper, the authors propose a pattern recognition method to detect journal IF manipulation. First of all, we define those journals as 'abnormal journals' which are suspected of manipulating their self-citations. The research procedure was as follows: collection of data on IFs and citation distributions of known manipulated journals in the *JCR* as a training set, extraction of the effective features for manipulated journals, construction of an original matrix model, and selection of *k*-nearest neighbours (*k*-NN) as the classifier to train samples. After analysing the test results, the validity of this method is verified by recognizing several unknown journals. Using the pattern recognition approach, abnormal journals can be distinguished in the *JCR* database.

2. Method

2.1. Pattern recognition

The term 'pattern recognition' (also known as machine identification, computer identification, or automatic machine identification) refers to information processing and analysis of various kinds of objects or phenomena.

Pattern recognition systems have four major components: data acquisition and collection, feature value extraction and representation, similarity detection and pattern classifier design, and performance evaluation [13]. Currently, the main techniques of pattern recognition are statistical pattern recognition, syntactic pattern recognition, fuzzy mathematics, neural networks, artificial intelligence methods, and data mining, or combinations of these. Statistical pattern recognition theory is still being perfected, and its methods are so numerous and effective that they now constitute an entire field of inquiry [14]. The basic technologies in this area are cluster classification, statistical decision-making methods, and nearest-neighbour classification. The nearest-neighbour method is the basic classifier in pattern recognition and is used in this paper.

2.2. k-nearest neighbour classification

The nearest-neighbour rule is a classification method based on the distance among samples, first proposed by Cover and Hart in 1967 [15]. The nearest-neighbour rule is one of the simplest and most important methods in pattern recognition [16]. The well-known *k*-NN algorithm is a lazy learning algorithm based on statistics, which has proven successful in many applications [17]. In the field of text categorization, *k*-NN is simpler, works better, and has better manoeuvrability on different data sets than other algorithms, such as naive Bayesian, support vector machine, linear least squares, and neural networks. Based on the vector space model (VSM), *k*-NN has proved to be one of the best classification methods.

In *k*-NN classification, the training dataset is used to classify each member of a ‘target’ dataset. The structure of the data is that there is a classification (categorical) variable of interest (‘manipulated,’ or ‘non-manipulated’, for example) and a number of additional predictor variables (citation age, IF, self-citation rate, etc.). Generally speaking, the algorithm proceeds as follows:

1. For each row (case) in the target dataset (the set to be classified), locate the *k* closest members (the *k*-NNs) of the training dataset. A Euclidean distance measure is used to calculate how close each member of the training set is to the target row that is being examined.
2. Examine the *k*-NNs – which classification (category) do most of them belong to? Assign this category to the row being examined.
3. Repeat this procedure for the remaining rows (cases) in the target set.
4. Additionally, XLMiner also lets the user select a maximum value for *k*, builds models in parallel on all values of *k* up to the maximum specified value, and performs scoring on the best of the resulting models.

For identification, the NN of each classification sample should first be found, and the training dataset should be determined in advance. If only one nearest sample point is chosen to decide the class of a new sample, the sample size is 1, and the method is called 1-NN; if two NN sample points are chosen to decide the class of a new sample, the sample size is 2, and the method is called 2-NN, and so forth; if the sample size is *k*, the method is called *k*-NN. To find the NNs for the *k* samples in the training dataset, it is necessary to calculate the distances between the unknown sample point and all samples in the training dataset. Then count the samples from that with the smallest distance until the *k*th sample; that distance is the minimum distance to the NN. If this minimum-distance region of the training set contains more samples of class 1 in the training set, an unknown sample with a smaller distance can be classified as class 1; if it contains more samples of class 2 in the training set, an unknown sample with a smaller distance can be classified as class 2. A better explanation of the *k*-NN process is provided in Figure 2.

Let ‘*’ represent points in class 1, ‘□’ represent points in class 2, and the distribution of the original training samples be as shown in Figure 2. The samples cannot be separated by a general linear function, so the *k*-NN method is used. When a new sample arrives, 2-NN is used to classify it. If the two nearest points of the new sample belong to class 1, the new sample point ‘Δ’ is considered to belong to class 1.

In this paper, the dataset of *JCR* citation distributions of known manipulated journals is used as the training dataset, and the characteristic parameters (also known as ‘predictor variables’) of the manipulated citation distributions are extracted; then a matrix model is constructed to identify the manipulated journals. The *k*-NN classification is used to train on the dataset of *JCR* citation distributions of known journals, to obtain test results, and to validate the model by identifying random samples from unknown journals.

2.3. Algorithm

The computation proceeds as follows: given a set of *n* samples which can be divided into two populations (*G1, G2*). Its characteristics can be described by *m* attributes. Before analysis, samples of each population can be characterized as:

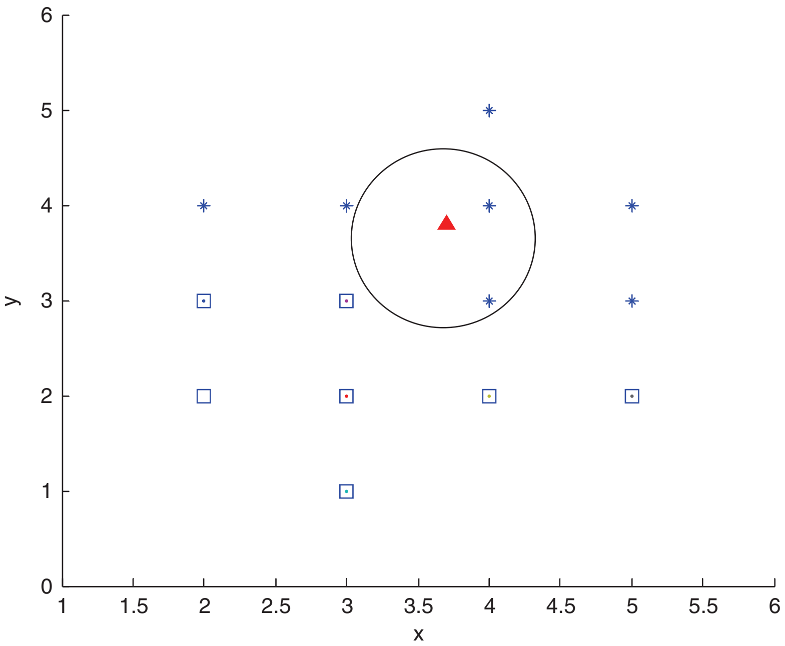

Figure 2. Nearest-neighbour classification results when *k* = 2 (*x*, *y* are non-dimensional)

$$\begin{matrix} X^1_{11} & X^1_{12} & \cdots & X^1_{1m} & X^2_{11} & X^2_{12} & \cdots & X^2_{1m} \\ X^1_{21} & X^1_{22} & \cdots & X^1_{2m} & X^2_{21} & X^2_{22} & \cdots & X^2_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ X^1_{n1} & X^1_{n2} & \cdots & X^1_{nm} & X^2_{n1} & X^2_{n2} & \cdots & X^2_{nm} \end{matrix} \tag{1}$$

(1) Construct an Euclidean distance function between an unknown sample X and an object Y belonging to class j:

$$D^2(X,Y)=(X-Y)^T\sum_j{}^{-1}(X-Y)\,, \tag{2}$$

where $(X - Y)^T$ is the neighbouring difference vector in the same index between object X and object Y of class j and $\sum_j^{-1}(X-Y)$ is the inverse of the covariance matrix which consists of the neighbouring differences in the same index between object X and object Y of class j.

(2) k-NN estimates of the density function of G_j and its *a priori* probability P_j can be expressed as:

$$f_j(x)=\frac{k_j}{nV_n(k,x)} \tag{3}$$

$$p_j=\frac{n_j}{n} \tag{4}$$

where k_j is the number of the selected k_n samples that belong to population G_j;

(3) Determine the posterior probability function:

$$P(G_i|\mathrm{X})=\frac{p_j f_j(x)}{\sum_{i=1}^{2} p_j f_j(x)} \tag{5}$$

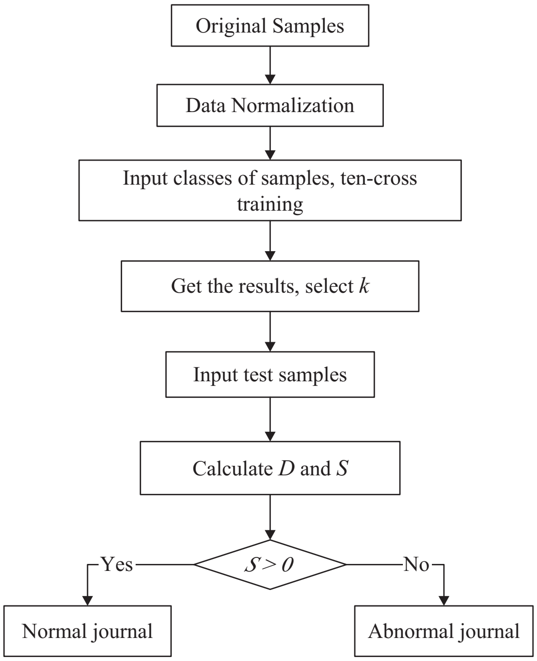

Figure 3. *k*-NN calculation procedure

(4) Construct a discrimination rule: if

$$p_i f_i(x) = \max_{1 \le j \le 2} p_j f_j(x)\ , j = 1, 2, \text{ then } x \in G_i.$$

This means that x belongs to the population with maximum posterior probability. If

$$p_1 f_1(x) = \max_{1 \le j \le 2} \frac{n_1}{n} \frac{k_j}{V_n(k,x)}\ , \text{ then } x \in G_i.$$

Therefore, the discrimination function also can be described as: if

$$k_i = \max_{1 \le j \le 2} k_j\ , \text{ when } i = 1, 2, \text{ then } x \in G_i.$$

It is known that the new sample belongs to the population which includes the greatest number of samples from among the k_n samples.

(5) The selection criterion is that the ability to distinguish be as good as possible. Discrimination results are generally evaluated by estimating the decision error rate and the decision accuracy rate. Low decision error rate and high decision accuracy rate mean good discrimination. This is achieved through judging the effect of training the samples and by selecting the appropriate k_n to achieve the purpose of the experiment. The resulting model is tested in practice to determine the sensitivity and reliability of the model and method.

k-NN is a simple pattern recognition classifier. It obtains classification results by collecting, training, and testing the original data. When a new sample is presented, the *k* nearest samples to the new sample in the training set are found. According to the labels of those *k* samples, the value for each class is calculated using the method of distance-weighted scores. Using the classification function, the label of the new sample can be obtained. When *k = 2*, the calculated discrimination function can be expressed as follows:

$$S = \sum_{i=1}^{k} \frac{S_i}{D_i^2} \tag{6}$$

where S_i is the class attribute value of sample i ($i = 1, 2$), the sample value of class 1 is positive, and the sample value of class 2 is negative. D_i^2 is calculated as in Equation 2. If the value of S is positive, the new sample belongs to class 1, otherwise it belongs to class 2. In this paper, normal journals are defined as class 1 and abnormal journals as class 2. The *k-NN* method is used to classify the original data and to obtain classification results from training on them. Then new data are collected and entered into the model, and the labels of the new samples are determined. The whole procedure as designed is shown in Figure 3.

3. Experimental

3.1. Processing data and extracting feature values

To improve recognition accuracy, large-sample data are needed. The following 10 normal journals were identified from the *JCR*:

(1) *Information Processing and Management;*
(2) *Journal of Information Science;*
(3) *Scientometrics;*
(4) *Journal of Materials Chemistry;*
(5) *Current Opinion in Solid State and Materials Science;*
(6) *Transactions of the Nonferrous Metals Society of China;*
(7) *Journal of Materials Processing Technology;*
(8) *Journal of the European Ceramic Society;*
(9) *Ceramics International;*
(10) *Carbon.*

In this paper, normal journals are considered as those which have not requested authors to cite their journals. According to explanatory statements and announcements made by some authors after submitting their works to certain journals, seven abnormal journals (‘manipulated journals’) were selected in which the editors had asked authors to cite papers published in their journals in the previous two years. To protect their identities, the journals are referred to as ‘Journal 1’, ‘Journal 2’, ‘Journal 3’ … ‘Journal 7’. The total number of citations, total number of self-citations, various IFs, distributions of citations, and distributions of self-citations of the journals (both abnormal and normal) from 1998 to 2007 were determined from the JCR database. Then the number of self-citations in the previous two years was calculated for every journal. To obtain the optimal feature vector, the immediacy index and citation half-life were also determined, and a decision tree was used to analyse quantitatively the degree of every feature vector. After testing the influence of different feature combinations on classification performance, the dimension of the vectors was minimized, and the effective feature values were extracted. The final feature space was: $X(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$, where

- x_1 = total number of citations;
- x_2 = total number of self-citations;
- x_3 = number of citations in this year;
- x_4 = number of self-citations in this year;
- x_5 = number of citations in the previous two years;
- x_6 = number of self-citations in the previous two years;
- x_7 = self-citation rate.

However, this set of original features of the sample data is small relative to the set of journals in the *JCR*. Therefore, further preprocessing of the data, called here a resampling process, is necessary. Because journal publishing is a continuous process and contains time delays, a new feature space was defined that extends each feature value to two consecutive years as a statistical unit. In this paper, the original data consist of 18 matrices with 10 rows and seven columns. Because this data set is not large enough, it is necessary to preprocess the data by resampling the original data.

The resampling process can be illustrated using Journal 1 as an example. The original matrix of Journal 1, shown in Table 1, is a matrix with 10 rows representing the years from 1998 to 2007. Its feature space consists of 10 rows and seven columns. The feature values are shown in Table 1: $x_1, x_2, x_3, x_4, x_5, x_6, x_7$. The vertical dimension shows the data change from 1998 to 2007. After resampling, each feature value is extended over two consecutive years to reflect changes over time in the journals. As shown in Table 2, the feature space has been extended into a matrix with nine rows and 14 columns, containing 14 feature values. The first seven values are $x_1, x_2, x_3, x_4, x_5, x_6, x_7$, representing the first year, and now denoted as $N_1, N_2, N_3, N_4, N_5, N_6, N_7$; the last seven values are $x_1, x_2, x_3, x_4, x_5, x_6, x_7$, representing the consecutive second year, and now denoted as $M_1, M_2, M_3, M_4, M_5, M_6, M_7$. It is evident that, after resampling, this dataset is much more sufficient and can improve pattern recognition ability. Moreover, the feature values of different samples from the same class are similar, but the feature values of samples from different classes are different.

3.2. Experimental procedure and analysis of results

For dealing with massive data, programs from the MATLAB toolbox are called to achieve automatic operation. The optimal *k* value for the data model is obtained and analysed to determine whether it meets the requirements. The process is as follows:

Table 1. Original data matrix for Journal 1

Journal 1	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1998	392	62	0	0	37	6	0.158
1999	385	47	0	0	75	15	0.122
2000	474	46	0	0	64	3	0.097
2001	823	112	13	9	125	16	0.136
2002	816	109	9	1	119	21	0.134
2003	1000	100	13	1	99	18	0.100
2004	992	83	10	0	114	10	0.084
2005	1137	171	20	2	118	20	0.150
2006	1347	203	48	5	218	35	0.151
2007	1141	148	23	1	282	33	0.130

Table 2. Matrix for Journal 1 after resampling

Training sample	N_1	N_2	N_3	N_4	N_5	N_6	N_7	M_1	M_2	M_3	M_4	M_5	M_6	M_7
1	392	62	0	0	37	6	0.158	385	47	0	0	75	15	0.122
2	385	47	0	0	75	15	0.122	474	46	0	0	64	3	0.097
3	474	46	0	0	64	3	0.097	823	112	13	9	125	16	0.136
4	823	112	13	9	125	16	0.136	816	109	9	1	119	21	0.134
5	816	109	9	1	119	21	0.134	1000	100	13	1	99	18	0.100
6	1000	100	13	1	99	18	0.100	992	83	10	0	114	10	0.084
7	992	83	10	0	114	10	0.084	1137	171	20	2	118	20	0.150
8	1137	171	20	2	118	20	0.15	1347	203	48	5	218	35	0.151
9	1347	203	48	5	218	35	0.151	1141	148	23	1	282	33	0.130

Table 3. Results after normalization of Journal 1 data

Training sample	N_1	N_2	N_3	N_4	N_5	N_6	N_7
1	0.002	0.018	0.000	0.000	0.000	0.003	0.287
2	0.002	0.014	0.000	0.000	0.002	0.008	0.221
3	0.003	0.013	0.000	0.000	0.002	0.001	0.175
4	0.006	0.034	0.004	0.008	0.004	0.008	0.247
5	0.006	0.033	0.003	0.001	0.004	0.011	0.242
6	0.007	0.030	0.004	0.001	0.003	0.009	0.180
7	0.007	0.025	0.003	0.000	0.004	0.005	0.150
8	0.008	0.053	0.006	0.002	0.004	0.010	0.273
9	0.010	0.063	0.015	0.004	0.009	0.018	0.274

(1) Data preparation: the data matrices of all original feature values are integrated into one matrix. The result is a matrix with 162 rows and 14 columns. The first 90 rows of the integrated matrix contain feature values for normal journals, and the last 72 rows contain feature values for abnormal journals. The first seven columns represent the first year and the last seven columns the second year. Because the data set is so large that it cannot be shown here, only the new matrix for Journal 1 is presented in Table 2.

(2) Data processing: first, the original data are placed into the *Current Directory* window in MATLAB. A variable representing the original data, called *Data*, is then created in the *Workspace* window. The standardized program ‘fnnormal’ is then used to standardize *Data* and store it in a new variable, *Data normal*. Now the data for Journal 1 have been changed as shown in Table 3. Moreover, the label of each data point should be added into the last column of *Data normal* (the first 90 rows have a label of 1, corresponding to normal journals, while the last 72 rows have a label of 2, corresponding to abnormal journals). Finally, the coefficients of *k*-NN are chosen to be [10,3; 2,1]. The ‘*fnknn*’ and ‘*fncrossValidate*’ programs are then used for cross-validation.

(3) Ten-cross validation: the ‘*fncrossValidate*’ program is used to validate the accuracy of the model. The 10-cross validation process is complex. First, because the first parameter in ‘*coeff*’ is 10, *Data* normal is divided into 10 groups at random. In each experiment, a combination of nine groups is used as a training set, with the last group serving as a test set. Then the accuracy is calculated 10 times to obtain the average accuracy for all 10 groups. This

Table 4. Randomly selected groups

No	1	2	3	4	5	6	7	8	9	10	Total
Sample	17	16	16	16	16	16	16	16	17	16	162

Table 5. Accuracy based on 10 groups

No	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	0.941	1	1	1	1	1	1	1	1	1	0.994

average result is the accuracy rating of the model. For the present experiment, the random selection of 10 groups and the accuracy results are shown in Tables 4 and 5.

(4) Analysis of results: the recognition accuracy of this model can be expressed as:

$$\mathrm{R} = \frac{Q_2 - Q_1}{Q_2} \times 100\%$$

where R is the recognition accuracy, Q_1 is the total of the label numbers of journals which were falsely identified by 10-cross validation, and Q_2 is the total of the label numbers of all journals.
The label corresponds to the class of each journal. The label of a normal journal is 1, and the label of an abnormal journal is 2. The recognition results are also 1 or 2. The smaller the differences between the recognition results and the corresponding labels, the higher the accuracy of the model. In Table 5, the recognition results for 10 random groups are 94.1, 100, 100, 100, 100, 100, 100, 100, 100 and 100%. The average result for the model is 99.4%, which means that the model can recognize normal and abnormal journals with an accuracy of 99.4%.

Considering this process and the pattern recognition results, several conclusions can be drawn.

(1) In the present research, *k*-NN is a good classifier, achieving classification accuracy up to 99%. When noise is taken into account, the classifier still achieves a satisfactory result. This proves that the classifier is sensitive and reliable for the problems addressed.
(2) When processing massive data sets, MATLAB is a useful tool for pattern recognition and intelligent computing. Much time and effort were saved by using MATLAB programs to achieve a stable model and to identify manipulations.
(3) After learning and training on the original data, the authors decided to resample the feature values selecting two consecutive years of feature values as a unit. This approach yielded an effective model that could be applied to actual journals.

4. Verifying the method

Despite the high accuracy obtained from the model, it is still necessary to verify its validity. First, the pretreated data for the journals studied were entered into the model for validation. Then, using the *k*-NN program, a test label was obtained for each data point. The validation process is shown in Figure 4.

4.1. Recognition on known journals

The validity of the model was verified using the original data. According to the procedure shown in Figure 4, first of all, the feature value matrices of the original journals were entered into the model. The normal journal *Scientometrics* and an abnormal *Journal x* (known to be manipulated) were chosen as test journals. Then the model was normalized, and values of 1 or 2 were added in the last column of the new model. Now the test set is the set of feature values of normalized data for the test journal and the training set is the set of normalized data from the original model. After running the *k*-NN program, the label of the test journal will be displayed as a result. The labels obtained for the selected journals are shown in Tables 6 and 7.

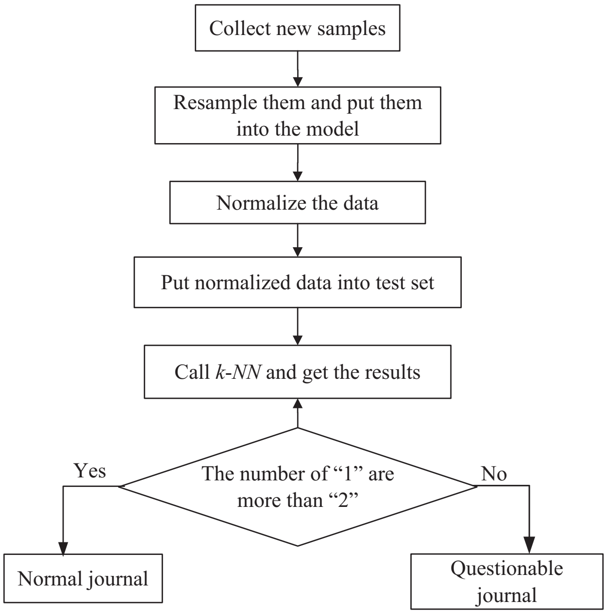
Collect new samples
Resample them and put them into the model
Normalize the data
Put normalized data into test set
Call k-NN and get the results
The number of "1" are more than "2"
Yes
No
Normal journal
Questionable journal

Figure 4. Validation process for known journals

Table 6. Recognition results for *Scientometrics*

Scientometrics test run	1	2	3	4	5	6	7	8	9
Label obtained	1	1	1	1	1	1	1	1	1

Table 7. Recognition results for *Journal x*

Journal x test run	1	2	3	4	5	6	7	8	9
Label obtained	2	2	2	2	2	2	2	2	2

Table 8. Feature values for the example *Journal y*

Example Journal y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
2003	113	94	53	51	58	42	0.832
2004	79	45	1	0	61	38	0.570
2005	168	71	16	8	84	34	0.423
2006	161	78	13	9	50	24	0.484
2007	216	115	7	7	63	35	0.532

It is clear that the labels of *Scientometrics* are all 1, which means that it is a normal journal with a probability of 100%. The labels of *Journal x* are all 2, which means that it is an abnormal journal. In other words, the results are consistent with the known facts.

4.2. Identifying unknown journals

To illustrate further the usefulness and generalizability of this method, one approach is to extract data on two journals from the *JCR* and to determine whether they are manipulated. The journals selected were new journals which can be easily manipulated. They may have high self-citation rates, and their IFs are less than 1. Then their feature values were extracted and entered into a matrix that was provided to the model. The feature values of the two new journals are shown in Tables 8 and 9. Thus, a new data sample *Data* was generated. According to the procedure, the data were then standardized into *Data Normal*. Then the label is added to the last column of the new model, and also a value of *coeff* was assigned.

Table 9. Feature values for the example *Journal z*

Example Journal z	x_1	x_2	x_3	x_4	x_5	x_6	x_7
2003	277	87	23	10	82	31	0.314
2004	211	46	5	2	78	21	0.218
2005	214	54	2	1	77	27	0.252
2006	204	62	16	7	55	12	0.304
2007	291	120	24	13	79	47	0.412

Table 10. Recognition results for *Example Journal y*

Example Journal y	1	2	3	4
Label	1	1	1	1

Table 11. Recognition results for *Example Journal z*

Example Journal z	1	2	3	4
Label obtained	2	2	2	1

The training set is the normalized original matrix of *Data Normal*, and the test set is the normalized data of the new journal. By running the *k*-NN program, it is possible to obtain recognition results for the new journal. For example, the data for *Example Journal y* were entered into the model and standardized. The recognition results (labels) obtained by running the program are shown in Table 10. These recognition results indicate that *Journal y* is a normal journal with a probability of 100%

Then the data for *Example Journal z* were entered into the model and standardized to obtain normalized data. The *k*-NN program was run, and the recognition results (labels) were obtained as shown in Table 11. The result can be expressed as a probability of 75%, indicating that *Example Journal z* is an abnormal journal with a very high probability. The accuracy of this result might possibly be improved with more data than just from 2003 to 2007. For now, the conclusion is that it is an abnormal journal. If a journal has a probability of around 50% of being abnormal, it may need to be checked to determine whether or not its papers have been manipulated in the previous several years.

5. Conclusions

The purpose of this paper is to identify journals for which the IFs have been manipulated. By selecting samples of normal and abnormal journals, the authors have constructed a recognition model using the *k*-NN method as a pattern recognition classifier. The feasibility of this method has been verified experimentally. From the recognition results obtained for journal IF manipulation, several conclusions can be drawn:

(1) Whether the model can identify manipulated journals depends on the sample size. Only a very large data set can properly reflect the features of the two types of journals (normal and questionable), and it is feasible to use pattern recognition techniques on such data. The larger the amount of data on journals that can be collected from the *JCR*, the more accurate will be the experimental results. Meanwhile, the extraction and reduction of feature values are related to the extent of discrimination between the two types of journals. The feature values should comply with some simple rules: that is, among the feature values, different samples from the same class should be similar, while samples from different classes should be very different.

(2) The dimension of the recognition space can be controlled using the *k*-NN method. To obtain optimal recognition results, try different feature-space dimensions to find an optimal space based on feature-value characteristics and then improve the pattern recognition accuracy rate. In the experiments, the average recognition accuracy of this model was as high as 99%. Therefore, recognizing journal IF manipulation using *k*-NN has been shown to be an effective approach.

(3) Furthermore, this method has been verified by recognizing several additional journals. From the validation results, it is clearly feasible to recognize manipulated journals using pattern recognition. This method can be used to identify abnormal journal behaviour. Use of pattern recognition can effectively control the appearance and spread of IF manipulation.

In practice, as in all pattern recognition results based on probabilities, small errors are inevitable. For journals that yield ambiguous recognition results, it is possible to determine their correct label manually by checking journal citation data. The journal development process has its own regularity. Scientific tools can be used to supervise behaviour of journal editors and to monitor journal development. This paper describes the design of a pattern recognition method, but one which still needs to be perfected in practice. The purpose of this work is to reveal the true features of science development and publication. The objective is to ensure that journals are developing along a normal path rather than being influenced by human factors.

Funding

This project was supported by the National Natural Science Foundation of China (grant no. 70973031) and the United Fund of ISTIC-Thomson Reuters (IT2010007).

References

[1] W. Kuo and J. Rupe, R-impact: reliability-based citation impact factor, *IEEE Transactions on Reliability* 56(3) (2007) 366–367.
[2] P. Dong, M. Loh and A. Mondry, The ‘impact factor’ revisited, *Biomedical Digital Libraries* 2(7) (2005).
[3] R. Smith, Journal accused of manipulating impact factor, *British Medical Journal* 314(15) (1997) 461.
[4] J.H. Fowler, D.W. Aksnes, Does self-citation pay?, *Scientometrics* 72(3) (2007) 427–437.
[5] D.W. Aksnes, A macro study of self-citation, *Scientometrics* 56(2) (2003) 235–246.
[6] H. Snyder, Patterns of self-citation across disciplines (1980–1989), *Journal of Information Science* 24(6) (1998) 431–435.
[7] A. Hemmingsson, Manipulation of impact factors by editors of scientific journals, *American Journal of Roentgenology* 178(3) (2002) 767.
[8] G. Yu and L. Wang, The self-cited rate of scientific journals and the manipulation of their impact factors, *Scientometrics* 73(3) (2007) 356–366.
[9] C. Wallner, Ban impact factor manipulation, *Science* 323 (2009) 461.
[10] M.E. Falagasi and V.G. Alexioui, The top ten in journal impact factor manipulation, *Ethics in Science* 56 (2008) 223–226.
[11] G. Yu, D.H. Yang and L. Wang, Reliability-based citation impact factor and the manipulation of impact factor, *Scientometrics* 83(1) (2010) 259–269.
[12] F.T. Krell, Should editors influence journal impact factors?, *Learned Publishing* 23(1) (2010) 59–62.
[13] A.Rosenfeld and H. Wechsler, Pattern recognition: historical perspective and future directions, *International Journal of Imaging Systems and Technology* 11(2) (2000) 101–116.
[14] A.K. Jain, R.P.W. Duin and J. Mao, Statistical pattern recognition: a review, *Pattern Analysis and Machine Intelligence* 22(1) (2000) 4–37.
[15] T. Cover and P. Hart, Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* 13(1) (1967) 21–27.
[16] W. Zuo, D.Zhang and K.Q. Wang, On kernel difference-weighted *k*-nearest neighbor classification, *Pattern Analysis & Applications* 11 (2008) 247–257.
[17] S.E. Buttrey and C. Karo, Using *k*-nearest-neighbor classification in the leaves of a tree, *Computational Statistics & Data Analysis* 40 (2002) 27–37.