This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2016.2612265, IEEE Transactions on Fuzzy Systems

Exploiting Data Reliability and Fuzzy Clustering for Journal Ranking

Pan Su[a], Changjing Shang[b], Tianhua Chen[b] and Qiang Shen[b]
[a]School of Control and Computer Engineering, North China Electric Power University, Baoding, China
[b]Department of Computer Science, Institute of Mathematics, Physics and Computer Science
Aberystwyth University, Aberystwyth SY23 3DB, UK

***Abstract*—Journal impact indicators are widely accepted as possible measurements of academic journal quality. However, much debate has recently surrounded their use, and alternative journal impact evaluation techniques are desirable. Aggregation of multiple indicators offers a promising method to produce a more robust ranking result, avoiding the possible bias caused by the use of a single impact indicator. In this paper, fuzzy aggregation and fuzzy clustering, especially the Ordered Weighted Averaging (OWA) operators are exploited to aggregate the quality scores of academic journals that are obtained from different impact indicators. Also, a novel method for linguistic term-based fuzzy cluster grouping is proposed to rank academic journals. The work allows for the construction of distinctive fuzzy clusters of academic journals on the basis of their performance with respect to different journal impact indicators, which may be subsequently combined via the use of the OWA operators. Journals are ranked in relation to their memberships in the resulting combined fuzzy clusters. In particular, the nearest-neighbour guided aggregation operators are adopted to characterise the reliability of the indicators, and the fuzzy clustering mechanism is utilised to enhance the interpretability of the underlying ranking procedure. The ranking results of academic journals from six subjects are systematically compared with the outlet ranking used by the Excellence in Research for Australia (ERA), demonstrating the significant potential of the proposed approach.**

***Index Terms*—Journal ranking, fuzzy clustering, aggregation of indicators, OWA, data reliability, ERA.**

I. Introduction

THE assessment of research output quality is a serious issue which relates to many educational and financial problems such as evaluation of research projects and distribution of research funding. Recently many countries have implemented their own national projects for academic output assessment. Examples include the Research Excellence Framework (REF) in the UK [1] and the Excellence in Research for Australia (ERA) [2]. One significant aspect of research quality assessment may involve academic journal ranking, though the efficacy of using such information is not universally agreed upon [3]. However, the rank of a journal typically implies its prestige, impact, and even difficulty of having a paper accepted for publication in it. Nevertheless, the general concept of academic journal quality is a multi-faceted notion. Conventionally, assessing the quality of research publications is done through subjective peer-review, which is carried out by experts in the relevant research areas. It is almost inevitable that such expert-based assessment is expensive and time consuming, despite the issue of subjecrtivity. For example, in the ERA, over 700 experts were employed to make a journal ranking list. Although the sophisticated results judged by the experts can be very useful in, for instance, directing government research funding and reflecting appropriate use of public funds, the running costs involved make it impracticable to implement such approaches frequently.

The most recent methods for the ranking of academic journals rely on developments in computer and information technologies. Many on-line academic publication databases allow for access to not just the journals themselves, but the statistical information regarding their impact. The impact of academic journals is typically gauged using metrics such as the *Thomson Reuters Impact Factor* (IF) [4] (which is arguably the best known and most used), the 5-year IF [5], the Eigenfactor [6] and the SCImago Journal Rank [7]. A number of these factors have been successfully applied in creating the popular *Thomson Reuters Journal Citation Report* (JCR) [8], which provides a quantitative tool for the ranking, evaluation and comparison of academic journals to be carried out in a potentially objective manner. However, each indicator has its own strengths and limitations, and the results of their use can be quite diverse [9] and should be considered with due caution.

Recent trends for the evaluation of the impact of academic journals focus on utilising advanced computational methods rather than pure statistical indicators. For instance, the work of [10] examines the publishing behaviour of full-time, tenured faculty members from leading universities in order to rate journals (in the carefully selected area of information science research). Such a behaviour-based approach assumes that the collective publication record of research members at a sizable set of leading research universities is representative of good journals that make the greatest contributions to the research field concerned. Also, in the work on Reader Generated Networks (RGN) [11], inter-connections amongst journals are captured on the basis of the download sequence of their publications, extracted from a digital library download log. The journal impact rankings are then calculated from the resulting networks using various social networking centrality metrics. Empirically, the indicators derived from an RGN reflect different views from conventional journal impact evaluation, and its final ranking list may significantly deviate from that which is obtained by the direct use of IF.

In general, although much debate surrounds the (over-)use of journal impact indicators, especially in their individual forms (for example, the impact factor may be subject to manipulation by interested parties, including publishers or editors), they are nonetheless widely accepted by scholars as an objective tool to balance the drawbacks of pure human peer-review. As pointed out in [12]: "*the solution appears to be in a combination of peer review and objective indicators. These indicators should be assessed for relevance and reliability*".

An intuitive way to improve the reliability of indicators is the integration of multiple metrics. In the literature, Choquet integral classifiers [13] have been employed to integrate different indicator scores which are reported in the JCR in an effort to predict the journal ranks published by the ERA 2010 [14]. Another approach is to fuse various indicator scores through the use of distance/similarity measures [15], in which journals are placed in a multi-dimensional space and each dimension reflects a certain impact indicator. In [16], a statistical model is proposed to cluster journals which are collected from the Italian research evaluation exercise over the period of 2004-2010. It exploits an extended latent class model (for polytomous item response data) to estimate the grades of journals and subsequently to cluster and rank them. Whilst promising, there is much to be done in making these techniques more robust and generic in order to support activities such as the aforementioned research quality assessment.

The interpretability of the existing numeric indicators also raises a practical issue. This is because direct use of precise numerical values makes it difficult to integrate such objective indicators and peer review results, with the latter typically being given in qualitative terms. To enhance both the relevance and reliability of numeric indicators, this paper proposes an approach for building aggregated fuzzy clusters between journals based on their indicator scores. For each individual indicator, fuzzy clusters of journals which are labelled with qualitative linguistic terms are generated. The OWA operators are then employed to aggregate various fuzzy clusters according to their linguistic labels, thereby constructing the final fuzzy clustering results. Further, two refinement methods are introduced in order to derive a ranking of journals according to their memberships in the resulting fuzzy clusters. The overall ranking process is not only more reliable and interpretable than ranking by the original indicator scores but also very intuitive. The proposed approach is tested on six datasets of journals representing different academic subject areas and the ranking results are compared with those given by human experts. Experimental results demonstrate that the techniques developed in this work help to reflect and assess the impact of academic journals effectively.

The remainder of this paper is organised as follows. Section II introduces the basics of the OWA aggregation operators and fuzzy clustering. Section III describes the details of the proposed fuzzy aggregation and cluster grouping for journal ranking. Section IV presents the experimental evaluation of the proposed approach, along with a discussion of the empirical results. Finally, Section V concludes the paper, including suggestions for further development.

II. Preliminaries

A. OWA Aggregation

When dealing with real-world problems, the opinions of different experts are usually aggregated in order to provide more robust solutions. Similarly, numeric measures of certain properties are also typically aggregated when addressing a given problem, with the weighted average operator being popularly adopted to implement the aggregation process [18], [19]. Apart from classical aggregation mechanism (such as average, maximum and minimum), another interesting and more general type of aggregation operator is the family of Ordered Weighted Averaging (OWA) operators [20], [21], [22]. OWA is a parameterised operator based on the ordering of extraneous variables that it is applied to. The fundamental aspect of this family of operators is the reordering step in which the extraneous variables are rearranged in descending order, with their values subsequently integrated into a single aggregated one.

Formally, a mapping $A_{\text{owa}} : \mathbb{R}^m \rightarrow \mathbb{R}$ is called an OWA operator if

$$A_{\text{owa}}(a_1, \cdots, a_m) = \sum_{i=1}^{m} w_i a_{\pi(i)} \tag{1}$$

where $a_{\pi(i)}$ is a permutation of a_i, which satisfies that $a_{\pi(i)}$ is the i-th largest of the a_i, and $w_i \in [0, 1]$ is a collection of weights that satisfies $\sum_i w_i = 1$, $i = 1, \cdots, m, m > 1$.

Without causing notational confusion, for simplicity, both the variables and their values are herein denoted as a_i, and are simply termed arguments. Also, the weights of an OWA operator are hereafter denoted as a weighting vector $W = (w_1, \cdots, w_m)$, in which the i-th component is w_i. Different choices of the weighting vector W can lead to different aggregation results. The ordering of input arguments gives OWA a nonlinear feature. Three special cases of the OWA operator are the classical *mean*, *max* and *min*. The *mean* operator results by setting $w_i = 1/m$, the *max* by $w_1 = 1$ and $w_i = 0$ for $i \neq 1$, and the *min* by $w_m = 1$ and $w_i = 0$ for $i \neq m$. These weighting vectors are denoted as W_{mean}, W_{max} and W_{min} respectively, in the remainder of the paper. Obviously, an important feature of the OWA operator is that it is a mean operator which satisfies

$$\min\{a_1, \cdots, a_m\} \leq \sum_{i=1}^{m} w_i a_{\pi(i)} \leq \max\{a_1, \cdots, a_m\}. \tag{2}$$

Such an operator provides aggregation between the maximum and the minimum of the arguments. This boundedness implies that it is idempotent; that is, if all $a_i = a$ then $A(a_1, \cdots, a_m) = a$.

As such, different weighting vectors can be devised in order to express different aggregation behaviours of the OWA used. A measure which is commonly employed to reflect the overall behaviour of an OWA operator is orness [23]. It captures the essential design intention of whether an aggregation operator will behave similarly to the interpretation of logical conjunction (influenced by smaller inputs) or that of disjunction (influenced by larger inputs). In particular, an orness measure of an OWA operator with the weighting vector W (named attitudinal character [20]) is defined by

$$\text{orness}(W) = \frac{1}{m-1}\sum_{i=1}^{m}((m-i)w_i). \tag{3}$$

The higher the orness value, the more similar the aggregated result is to that of disjunction. In particular, it can be calculated that $\text{orness}(W_{mean}) = 0.5$, $\text{orness}(W_{max}) = 1$ and $\text{orness}(W_{min}) = 0$.

B. OWA Aggregation with Dependent Weights

When combining multiple arguments using a pre-defined weighting vector in OWA, the weights in aggregation are normally assumed to be argument-independent as they are not necessarily related to the extraneous variables they are applied to. Therefore, the use of unduly high or low weights should be avoided. Otherwise, a typical OWA operator may suffer from giving the highest priority to outlier variable values [24], leading to the generation of false or biased judgments when the operator is in action. To achieve more reliable outcomes, a type of OWA operator with dependent weights has been introduced in the literature, in which the normal-distribution of argument values is used to determine the weight vector. This type of OWA operator helps reduce the risk of obtaining biased results due to extreme outliers in the given extraneous variables.

In particular, the Dependent OWA (DOWA) operators [25] utilise weighting vectors that are derived in accordance with the average of arguments. Let $(a_1, a_2, \cdots, a_m)$ be the argument vector, and $\mu = \frac{1}{m}\sum_{i=1}^{m} a_i$. The similarity between any argument a_i and the average value μ can be calculated as follows:

$$s(a_i, \mu) = 1 - \frac{|a_i - \mu|}{\sum_{j=1}^{m}|a_j - \mu|}. \tag{4}$$

From this, a weighing vector can be generated by applying the following:

$$w_i = \frac{s(a_i, \mu)}{\sum_{j=1}^{m} s(a_j, \mu)} \tag{5}$$

$$A_{\text{dowa}}(a_1, \cdots, a_m) = \sum_{i=1}^{m} w_i a_i \tag{6}$$

Apart from measuring the reliability of arguments by their distances to the average value, there are alternative approaches. In kNN-DOWA [26] for example, the reliability of an argument is determined by its nearest neighbours. This type of reliability helps differentiate amongst a collection of arguments such that an argument whose value is similar to its k neighbours [27] is deemed reliable and can be assigned a high weight. In contrast, an argument that is largely different from its neighbours is discriminated as an unreliable member. Formally, the reliability measure of an argument a_i, $i = 1, \cdots, m$ in kNN-DOWA is defined as:

$$R_i^k = 1 - \frac{\sum\limits_{t=1}^{k} d(a_i, n_t^{a_i})/k}{\max\limits_{j,j' \in \{1,\cdots,m\}} d(a_j, a_{j'})} \tag{7}$$

where $n_t^{a_i}$ is the value of t-th nearest neighbour ($t = 1, \cdots, k$) of the argument a_i, and the distance measure d used to perform neighbour-searching is $d(a_j, a_{j'}) = |a_j - a_{j'}|$. This absolute distance metric is adopted for computational simplicity, but any other distance metric may be employed also.

Having obtained the reliability values of all arguments concerned, they are normalised to form the weighing vectors in kNN-DOWA. Given the reliability value R_i^k of each argument a_i, the corresponding kNN-DOWA operator $A_{\text{dowa}}^k : \mathbb{R}^m \to \mathbb{R}$ can be specified by

$$A_{\text{dowa}}^k(a_1, \cdots, a_m) = \sum_{i=1}^{m} w_i^k a_i \tag{8}$$

where $w_i^k = R_i^k / \sum_{j=1}^{m} R_j^k$. kNN-DOWA and DOWA are order independent (termed *neat* in the literature) [28], as they generate the same outcome regardless of the order of argument values. Another development, but of similar principle to kNN-DOWA, is the work of Cluster-DOWA where clusters of arguments are exploited to detect outliers in order to improve data reliability [24]. The common assumption made in all these methods is that arguments which have high reliability values should be weighted highly.

C. Fuzzy Clustering

Clustering is one of the important approaches within the framework of unsupervised learning which aims to assign objects into groups (namely clusters) such that objects in the same group are similar to each other, and dissimilar to those in the other clusters [29]. If a crisp clustering algorithm such as k-means [30] is used in the generation of clusters, the association degree of a data point belonging to a specific cluster is either 1 or 0. However, there are other popular clustering algorithms such as fuzzy c-means [31] that naturally produces clusters of data with uncertain boundaries. Fuzzy c-means is effective in generating fuzzy partitions for a given dataset. Each cluster in a fuzzy partition $\widetilde{\pi}$ is a fuzzy set $\widetilde{C}_k, k = 1, \cdots, K$ where $\widetilde{C}_k(x) \in [0, 1]$ represents the degree of a data point $x \in X$ belonging to the corresponding fuzzy cluster. Usually, this degree is normalised with all the clusters in a partition satisfying $\sum_{k=1}^{K} \widetilde{C}_k(x) = 1$.

Note that the key difference between a crisp clustering and a fuzzy one is that the latter produces fuzzy clusters. If the fuzzy clusters are defuzzified into crisp clusters, many techniques working on crisp clusters may be directly used for handling fuzzy clusters. Unfortunately, in so doing, invaluable information may be lost during the defuzzification process and therefore, the quality of the results may be adversely affected [32], [33]. Besides, the interpretability of the fuzzy linguistic terms inherent to the fuzzy approach would be missed. Owing to the non-binary memberships, useful information such as k-nearest neighbours can be extracted. To reflect this observation the present work proposes a fuzzy aggregation based method for aggregating fuzzy clusters, which is tailored for ranking academic journals.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2016.2612265, IEEE Transactions on Fuzzy Systems

III. Fuzzy Aggregation and Clustering based Journal Ranking

A. *Outline of the Approach*

With the aid of on-line academic publication databases such as IEEE Xplore, Scopus, and DBLP, the calculation of individual journal impact indicators can be carried out effectively. A number of indicators are widely accepted and applied by scholars, which typically aim to evaluate a single journal or work on one particular aspect of journal citations. Scores of journals gained from various indicators can be directly aggregated by using the aggregation methods mentioned previously. However, when human experts assess the quality of academic journals, linguistic terms are commonly and sensibly used to support their judgement. Therefore, the interpretable estimation of journal quality with respect to labelled fuzzy clusters (rather than the numerical scores) is utilised in this paper in order to perform journal ranking.

In addition to the classical IF and Eigenfactor metrics, other indicators of different focusses can also be employed. However, none of these is diverse enough to be able to individually characterise all aspects of journal impact by itself in the real-world. To compensate for the potential bias of using single indicators, thereby enhancing the reliability and relevance of memberships of journals to those labelled fuzzy clusters, a linguistic term based integration (i.e., consensus) method is proposed here, to regroup the fuzzy clusters generated by different indicators. Also, OWA operators with dependent weights are applied to implement the integration of fuzzy memberships.

The proposed journal ranking method is named FAC to reflect the fact that it is based on *fuzzy aggregation* and fuzzy *clustering*. Briefly, its working process starts by creating fuzzy clusters, using fuzzy c-means individually on each of the journal impact indicators which are available (and selected) from databases of academic publications. The resultant (fixed number of) fuzzy clusters, termed base clusters for easy reference, are associated with predefined linguistic labels. The preference relation amongst linguistic terms is then employed to group the base clusters. The OWA operators are used to aggregate the memberships of base clusters belonging to the same group, forming the final fuzzy clusters. The method may also involve the following two optional steps: 1) defuzzifying the resultant fuzzy clusters such that each data point (i.e., journal) belongs to just one final crisp cluster (which may still be associated with a linguistic label) and hence, introducing a relative ranking amongst all journals; and 2) combining the memberships of a given journal from all fuzzy clusters into a single index of rank, thereby giving an absolute rank amongst all journals. An illustrative flowchart of the FAC algorithm is shown in Fig. 1 and the following subsections detail its key operations.

B. *Indicator-based Generation of Fuzzy Clusters*

In translating a set of real-valued scores into a linguistic term which is closer to the use of natural language, it is a common practice to employ fuzzification techniques. For the present work, fuzzy c-means, which is able to retain the non-binary memberships of each data point in all clusters, is adopted to translate the numerical indicator scores into predefined linguistic terms. Without losing generality, suppose that a set of journals J is evaluated with regard to $m, m > 1$ impact indicators $I_1, \cdots, I_m$, and that each indicator I_h is a mapping $I_h : J \to \mathbb{R}, h = 1, \cdots, m$. Also, it is intuitively presumed that a higher impact indicator score is assigned to a journal with higher impact. For each indicator I_h, fuzzy c-means is then utilised to form clusters in J with respect to $\{I_h(j) | j \in J\}$ and a pre-specified number K (which indicates the number of fuzzy subsets in J that are required to be constructed). From this, K fuzzy sets are formed with $\widetilde{C}_1^h(j), \cdots, \widetilde{C}_K^h(j)$ representing the memberships of a journal $j \in J$ belonging to the resulting individual fuzzy clusters, respectively.

When linguistic terms are employed to describe a variable, a preference ordering relation is usually defined on the set of linguistic terms such as $Bad \prec Acceptable \prec Good$ or $Low \prec Medium \prec High$. In the general application of fuzzy clustering, such an ordered labelling scheme over the clusters is not necessary. However, in FAC, labelling the clusters is not only helpful to understand the relative quality of journals in a cluster, but also important to organise base clusters in the subsequent aggregation process. The required labelling may be accomplished by consulting human experts in the field. Yet, since the fuzzy clusters are herein generated according to a given individual impact indicator whose values are totally ordered, the value of each cluster centre can be employed to signify the overall relative quality of that cluster. Thus, given a set of $K (K > 1)$ pre-defined linguistic terms $L = \{L_1, \cdots, L_K\}$ which satisfy that $L_1 \prec \cdots \prec L_K$, the fuzzy clusters $\widetilde{C}_1^h, \cdots, \widetilde{C}_K^h$ can be readily sorted in ascending order with regard to their cluster centres and then, are labelled with $L_1, \cdots, L_K$ respectively.

A possible drawback of employing fuzzy c-means to implement fuzzification is that a data point's membership to a cluster is not monotonically decreasing with its distance to the cluster centre. This is caused by the mechanism of normalisation which is inherent in the fuzzy c-means algorithm. If the fuzzy clusters are defuzzified into crisp clusters by assigning each object to the cluster with which it has the maximum membership, the non-maximum (and non-monotonic) memberships will have no impact upon the final crisp result and hence, will be ignored. However, in FAC, memberships of a journal to all those linguistically labelled clusters are useful in the subsequent aggregation. Therefore, a filtering precess is applied to the resultant fuzzy memberships to ensure that the membership of a journal to a cluster is monotonically deceasing with its distance to the cluster centre. Such a filtering process can be implemented using the following two steps:

1) For each labelled fuzzy cluster $\widetilde{C}_{L_k}^h, k = 2, \cdots, K$, set membership $\widetilde{C}_{L_k}^h(j) = 0$ for each $j \in J$ where $I_h(j)$ is smaller than the centre of $\widetilde{C}_{L_{k-1}}^h$; and for each labelled fuzzy cluster $\widetilde{C}_{L_k}^h, k = 1, \cdots, K-1$, set membership $\widetilde{C}_{L_k}^h(j) = 0$ for each $j \in J$ where $I_h(j)$ is greater than the centre of $\widetilde{C}_{L_{k+1}}^h$;

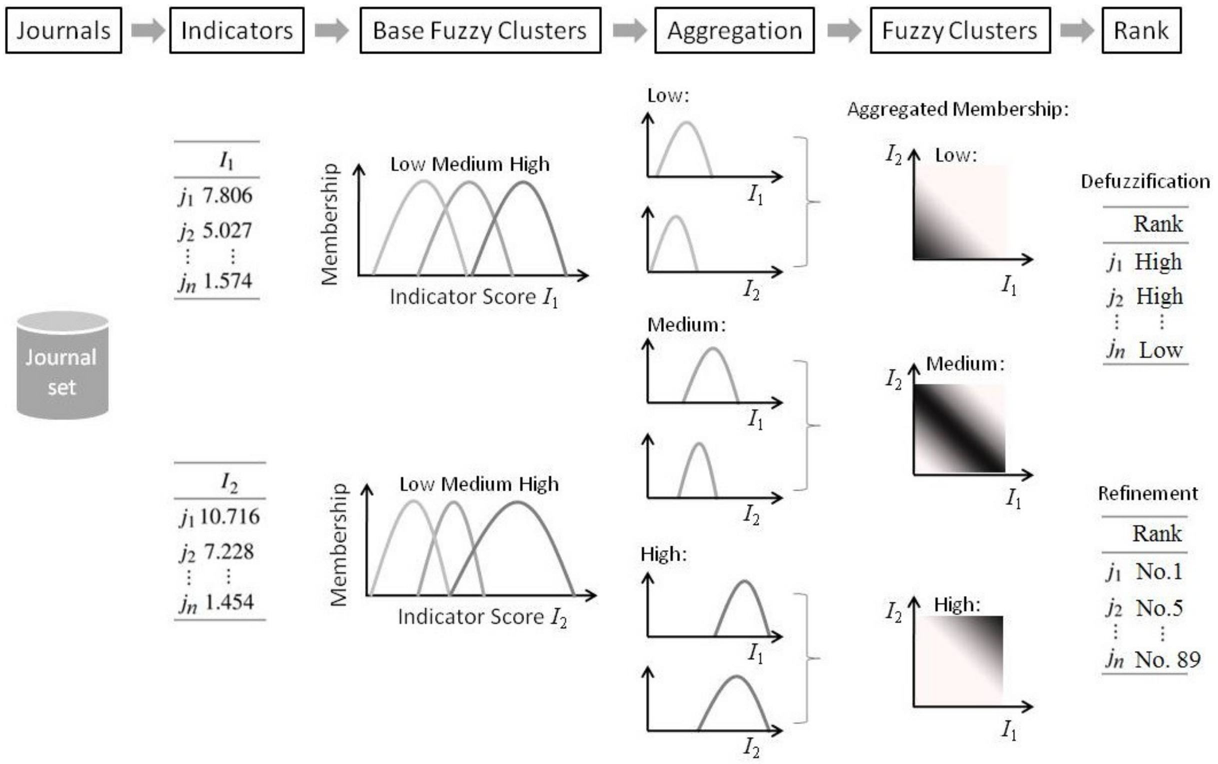

Fig. 1. FAC: Fuzzy Aggregation and Clustering based Journal Ranking

2) For each journal $j \in J$, update its memberships to all the clusters by normalisation:

$$\widetilde{C}^h_{L_k}(j) = \frac{\widetilde{C}^h_{L_k}(j)}{\sum_{i=1}^{K} \widetilde{C}^h_{L_i}(j)}. \tag{9}$$

Figure 2 shows an example of the above filtering precess. Fig. 2(a) is the fuzzy c-means result on a selective set of journals in Computer Science which are evaluated with their 2010 Impact Factor values. Fig. 2(b) is the filtered result using this method.

Note that Otsu [34] is a common method employed for one dimensional clustering in many applications such as image segmentation. However, fuzzy c-means is adopted within the present work instead of Otsu because it supports richer representation through the use of fuzzy memberships in which each data point may belong to several clusters (to better reflect the imprecision issue facing the current application problem). This allows more subtle information to be utilised in the aggregation of base-clusters. Note also that instead of performing clustering with respect to all indicators simultaneously (across all given data), the proposed approach aims to initially obtain a number of base clusters with regard to each individual indicator. This entails that the weights to be used for aggregating the effects of individual indicators can be assigned or learned through different means. In addition, it degenerates a task that otherwise requires simultaneous multi-objective optimisation to problems of single objective optimisation. Furthermore, the resulting clusters are less difficult to be labelled than those relying on the analysis of all indicators together without consulting experts, thereby minimising human intervention in the learning process.

Traditional statistical work on finite mixture models (FMM) and the expectation maximization (EM) algorithm can also be utilised to generate "soft" clusterings. It has been shown in [17] that partitions found by statistical or probabilistic approaches may be similar to those produced by fuzzy c-means on certain datasets. However, no matter what clustering algorithms are employed, even if a dataset is well separated into several recognisable subsets, the clustering methods may not always discover such structures because the otherwise appropriate parameters that could lead to a successful interpretation of the data are never used. As indicated previously, FMM has recently been employed to rank journals [16]. Such work can help determine the possibility of an instance belonging to a certain cluster, assuming that the distribution of data follows a certain statistical format. In particular, it provides more modelling flexibility than fuzzy c-means, by offering adaptable statistical parameters in the model. Nevertheless, choosing appropriate parameters for real-world data that will satisfy various modelling assumptions (e.g., polytomous item response data) can be a challenging task. By contrast, fuzzy c-means employs empirical heuristics and distance measures that work on imprecisely described domain values, not only facilitating the relaxation of otherwise required crisp discretisation, but also helping increase the interpretability of the clustering results. As it can be seen from the following subsections, the proposed approach reflects a bottom-up modelling strategy in obtaining journal clusters constructively, this differs from FMM which works by making

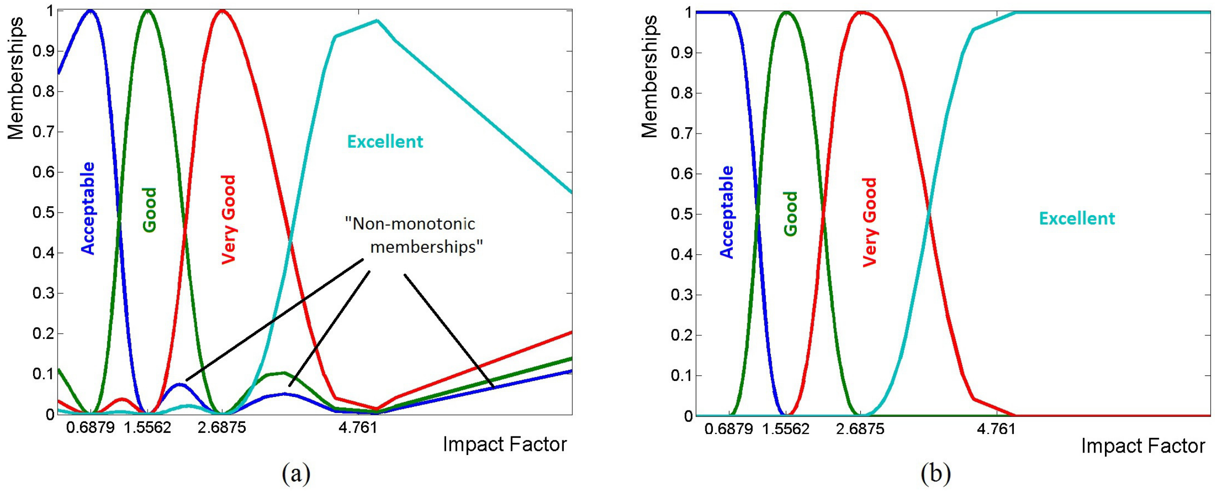

Fig. 2. Filtering the Fuzzy c-means Result

assumptions about the statistical model first and subsequently using the observed data to fit the model in a top-down manner.

C. Base Cluster Grouping and Aggregation of Memberships

Having gone through the fuzzification process as described in the preceding subsection, $m \times K$ fuzzy clusters are generated and labelled. In this step, the $m \times K$ fuzzy base clusters are grouped into K final clusters which are again labelled by the pre-defined set of linguistic terms L. Generally, this process can be seen as the consensus step in cluster ensemble [32]. However, cluster ensemble algorithms involve unsupervised grouping of base clusters, with many working methods available in the literature to implement such grouping, including: feature-based, graph-based, and voting-based, etc [35]. Since the fuzzy clusters of journals generated in FAC are automatically labelled (though using predefined linguistic terms), an intuitive "supervised" grouping of them becomes feasible (which is less challenging in implementation than using the unsupervised methods). This is described below.

Given the $m \times K$ labelled base clusters $\widetilde{C}^1_{L_1}, \cdots, \widetilde{C}^1_{L_K}$, $\widetilde{C}^2_{L_1}, \cdots, \widetilde{C}^{m-1}_{L_K}$, $\widetilde{C}^m_{L_1}, \cdots, \widetilde{C}^m_{L_K}$, owing to their inherent ordering, they can be (re-)categorised into K groups $C_1 = \{\widetilde{C}^h_{L_k} | k = 1, h = 1, \cdots, m\}, \cdots, C_K = \{\widetilde{C}^h_{L_k} | k = K, h = 1, \cdots, m\}$, where $\widetilde{C}^h_{L_k}$ is the fuzzy cluster which is generated by the impact factor I_h and labelled with L_k. C_k is a set of clusters, which contains all the clusters with the label $L_k, k = 1, \cdots, K$.

To illustrate the construction of C_k, for simplicity, a crisp counterpart of C_K is addressed first. Consider a voting system in which each indicator votes for the top-rated "excellent" journals, for example. Those in $C^h_{L_K}$ (i.e., the crisp counterpart of $\widetilde{C}^h_{L_K}$) are the journals voted by I_h and hence, C_K contains all the journals that are each regarded as an "excellent journal" by at least one of $I_1, \cdots, I_m$. Similarly, in general, C_k contains all those journals in the vote which are deemed to be of the quality level expressed by L_k. In such a crisp voting system, the votes can be summed for each journal and the winners can be ranked by how many ballots they have attracted. In FAC, however, each journal is not necessarily voted to have just one single quality level in a boolean way, but can have multiple explicit partial memberships assigned, indicating that it may be of different quality levels (though to various degrees). To make the best use of such information contained within such a voting system, more advanced aggregation operators rather than the simple sum/average are employed here to summarise the (both full and partial) votes, thereby deriving the final membership of a journal to a certain labelled fuzzy cluster C_k.

From this, given the K groups $C_1 = \{\widetilde{C}^h_{L_k} | k = 1, h = 1, \cdots, m\}, \cdots, C_K = \{\widetilde{C}^h_{L_k} | k = K, h = 1, \cdots, m\}$, the membership of a journal $j (j \in J)$ to the final, labelled fuzzy cluster $\widetilde{C}^*_{L_k}, k = 1, \cdots, K$ can be computed by:

$$\widetilde{C}^*_{L_k}(j) = A(\widetilde{C}^1_{L_k}(j), \cdots, \widetilde{C}^m_{L_k}(j)). \tag{10}$$

where A is an aggregation operator. Then, $\widetilde{C}^*_{L_k}$ is normalised by $\widetilde{C}^*_{L_k}(j) = \widetilde{C}^*_{L_k}(j) / \sum_{l=1}^K \widetilde{C}^*_{L_l}(j)$. The full algorithm of FAC is shown in Algorithm 1.

What is required now is the choice of a method to implement the aggregation operator A. As one of the possible mechanisms to perform the task of information aggregation, the concept of data reliability has been introduced [24], with successful extended applications for classification and feature selection. It works by exploiting the proximity to clusters of arguments and hence, can be rather inefficient. Recently, an enhanced version, termed kNN-DOWA, has been proposed in [26], where a hierarchical clustering process required by the original approach is replaced by a search of nearest neighbours. Whilst a number of aggregation operators are available in the literature and many of them have been applied to decision making [36], [37], they typically require subjective specification of the aggregation weights. Here, kNN-DOWA is adopted to aggregate the memberships of journals voted by different impact indicators. This is feasible because: 1) the weights used in the aggregation are learned from the arguments

Algorithm 1 FAC

```
Inputs: $J = \{j_1, \cdots, j_i, \cdots, j_n\}$: a dataset of $n$ journals, where $j_i = (I_1(j_i), \cdots, I_h(j_i), \cdots, I_m(j_i)) \in \mathbb{R}^m$ and $I_h(j_i)$ is the score of journal $j_i$ evaluated by the impact indictor $I_h$; $L = \{L_1 \prec \cdots \prec L_k \prec \cdots \prec L_K\}$: a set of $K$ linguistic terms with a preference relation;
Outputs: $\{\widetilde{C}^*_{L_1}, \cdots, \widetilde{C}^*_{L_K}\}$: $K$ labelled fuzzy clusters over $J$;

for $h = 1 : m$ do
    create sub-dataset $J_h = \{I_h(j_1), \cdots, I_h(j_n)\}$
    create base clusters $\widetilde{\pi}_h = \{\widetilde{C}^h_1, \cdots, \widetilde{C}^h_K\}$ using fuzzy $c$-means on $J_h$
    sort $\widetilde{C}^h_1, \cdots, \widetilde{C}^h_K$ to $\widetilde{C}^h_{\pi(1)}, \cdots, \widetilde{C}^h_{\pi(K)}$ so that the cluster center of $\widetilde{C}^h_{\pi(k)}$ is smaller than the cluster center of $\widetilde{C}^h_{\pi(k')}$, for $k < k'$ $(k, k' = 1, \cdots, K)$
    label $\widetilde{C}^h_{\pi(1)}, \cdots, \widetilde{C}^h_{\pi(K)}$ with $L_1, \cdots, L_K$ respectively, and gain $\widetilde{C}^h_{L_1}, \cdots, \widetilde{C}^h_{L_K}$
end for
regroup all the fuzzy clusters $\bigcup_{h=1}^m \widetilde{\pi}_h$ to create $K$ groups of fuzzy base clusters $C_1 = \{\widetilde{C}^1_{L_1}, \cdots, \widetilde{C}^m_{L_1}\}, \cdots, C_K = \{\widetilde{C}^1_{L_K}, \cdots, \widetilde{C}^m_{L_K}\}$
for $k = 1 : K$ do
    for $i = 1 : n$ do
        $\widetilde{C}'_{L_k}(j_i) = A(\widetilde{C}^1_{L_k}(j_i), \cdots, \widetilde{C}^m_{L_k}(j_i))$ where $A$ is an aggregation operator
    end for
end for
for $k = 1 : K$ do
    for $i = 1 : n$ do
        normalise $\widetilde{C}^*_{L_k}(j_i)$ by $\widetilde{C}^*_{L_k}(j_i) = \widetilde{C}'_{L_k}(j_i) / \sum_{l=1}^K \widetilde{C}'_{L_l}(j_i)$, such that $\sum_{k=1}^K \widetilde{C}^*_{L_k}(j_i) = 1$
    end for
end for
```

automatically; and 2) the weights assigned to the arguments represent their reliability, which can be collected as useful "by-products" to further analyse and interpret the reliability of the underlying impact indicators.

For a dataset with n points and m features, the time complexity of the original fuzzy c-means is $O(nmK)$, where K is the number of clusters [38]. Since FAC employs fuzzy c-means on a one dimensional dataset for m (the number of impact indicators) times, the time complexity of FAC in generating the base clusters is also $O(nmK)$. The time complexity of the consensus step depends on the aggregation operator A. Suppose that the complexity of aggregation is $O(A)$, then the overall time complexity of FAC is $O(nmK) + O(A)$. Take kNN-DOWA as an example, the time complexity of kNN-DOWA is $O(m^2)$ [26]. Therefore, if it is adopted to aggregate the memberships of journals, the complexity of the consensus step is $O(nm^2K)$, and the overall time complexity of FAC is $O(nmK) + O(nm^2K) = O(nm^2K)$.

D. Refinement for Ranking

Consider an example where the pre-defined set of linguistic terms is $\{Acceptable, Good, VeryGood, Excellent\}$ and the preference ordering relation is $Acceptable \prec Good \prec VeryGood \prec Excellent$. Suppose that the evaluation result of a journal using FAC is represented as a vector such as $(0.1, 0.1, 0.3, 0.5)$, whose elements denote the degrees of the journal belonging to the four (quality level) clusters, respectively. This form of result gives a "soft" evaluation of the quality of journals and is generally more informative than simply assigning journals to just one crisp cluster. Nevertheless, in many practical research quality assessment scenarios, it is not the absolute classification of journal qualities that is sought after, but the relative ranking amongst possible competitors. In order to decide on a rank of journals, using the information contained within the evaluation result vectors, two methods of transforming soft-partition to ranks are provided here.

The first is to assign a journal to the cluster(s) in which it has the maximum membership. That is, taking the strategy of the winner taking all. In so doing, the linguistic label associated with the final fuzzy cluster that possesses the maximum membership degree becomes the quality level of that journal, i.e.,

$$\text{rank of } j = \arg \max_{L_k \in L} \widetilde{C}^*_{L_k}(j). \tag{11}$$

Noted that L_K is the highest rank available for all journals while L_1 represents the lowest rank. Obviously, this method can only provide a fixed number of (i.e., K) ranks amongst the journals.

The alternative method is to assign a significance score to each of the linguistic terms and then, to sort the journals with respect to the weighted sum of the scores and journal (quality level) cluster memberships. For example, the scores can be set to $L_k = k$, reflecting the order of these quality levels. Then, the ranking over a set of journals can be obtained by sorting the journals in a descending order, according to:

$$\text{rank index of } j = \sum_{k=1}^K k \widetilde{C}^*_{L_k}(j). \tag{12}$$

Compared with the first method, this second approach can provide a more detailed ranking of the journals. The final ranks produced by the two methods are however, not necessarily in the same order. That is, journal j may be ranked higher than j' using the first method, but it may be ranked lower than j' if the second method is applied. The actual ranking outcomes depend on which method is used which in turn, depends on the results of the clustering. For example, suppose that the fuzzy evaluation of j is $(0.4, 0.0, 0.0, 0.6)$ and that of j' is $(0.0, 0.0, 0.6, 0.4)$, then j is ranked higher than j' using Eqn. (11) and lower using Eqn. (12). This is not a surprise, as these methods reflect different focuses of attention, similar to the use of conventional defuzzification techniques, where a different defuzzification method may result in a different defuzzified inference outcome. In a real application, so long as an approach is consistently utilised throughout, the ranking results will be consistent.

IV. Experimentation and Evaluation

A. Experimental Setup

The Journal Citation Report (JCR) [39] has a long history of applications for researchers and librarians in choosing their reading lists. All impact indicator score calculations in JCR are based on the same set of journals, namely journals which are indexed by the Web of Science. In order to test the performance of FAC, six indicators that are reported in JCR (2010) are selected as the indicators to construct base fuzzy clusters. These are [13]:

- Total Cites (TC): number of times the journal was cited in a year;
- Impact Factor (IF): ratio of cites to recent articles to the number of recent articles, with the recency being defined within a 2-year window;
- 5-year (5-IF): the same as IF, but covering articles within a 5-year window;
- Immediacy Index (II): ratio of cites to the current articles over the number of those articles;
- Eigenfactor (Ei): similar to IF, but eliminating self-referencing and weighting journals by the amount of time elapsed before being cited;
- Article Influence (AI): ratio of the Eigenfactor score to the total number of articles considered.

Generally, all these six indicators assign greater scores to journals with more citations. Apart from the indicators included in JCR, many other indicators are available from various of academic publication databases. Note that in this work, it is the methods that aggregate individual impact indicators together with data reliability that are investigated, rather than the selection of the underlying impact indicators themselves. Therefore, without losing generality, only the indicators reported in JCR are employed for testing here.

Note that these indicators have their own characteristics. As briefly defined above, Eigenfactor is developed to eliminate the effect of self-citation while IF and 5-IF include self-citation. AI is developed to offset the size effect of journals while TC does not take the size of a journal into consideration. However complex the interactions between these indicators may appear, they are more likely to be complementary to one another than to cause contradictions between each other. For example, an excellent journal can have high scores both in Eigenfactor and IF, and a journal which performs badly in TC may also perform badly in IF. In other cases, a journal could have higher scores in several indicators than in others. Due to the fact that they are proposed to measure journal quality with different focuses, direct comparison of the individual scores owing to their use can be difficult. After fuzzy clustering on each indicator, although the inherent characteristics of each indicator have not been changed, their numerical results have been mapped onto a new domain with interpretable linguistic meanings (e.g., *excellent*, *good*, etc). The resulting labelled clusters can therefore, help users to better understand the performance of journals from an integrated viewpoint of different perspectives.

In terms of datasets used for the experiments, six subject categories in JCR are selected, covering: *Chemistry* (Analytical, Applied, Inorganic & Nuclear, Medicinal, Multidisciplinary, Organic, Physical); *Computer Science* (Artificial Intelligence, Cybernetics, Hardware & Architecture, Information Systems, Interdisciplinary Applications, Software Engineering, Theories & Methods); *Materials Science* (Biomaterials, Ceramics, Characterisation & Testing, Coatings & Films, Composites, Multidisciplinary, Paper & Wood, Textiles); *Mathematics* (Applied, Interdisciplinary Applications); *Medicine* (General & Internal, Legal, Research & Experimental, Medical Ethics, Medical Informatics, Medical Laboratory Technology); and *Physics* (Applied, Atomic, Molecular & Chemical, Condensed Matter, Fluids & Plasmas, Mathematical, Multidisciplinary, Nuclear, Particles & Fields).

TABLE I
Summary of Datasets Used

		Number of Journals				
ID	Dataset	A*	A	B	C	Total
D1	Chemistry	37	70	95	143	345
D2	Computer Science	44	101	108	67	320
D3	Material Science	26	61	80	61	228
D4	Mathematics	52	84	127	69	332
D5	Medicine	20	39	73	107	239
D6	Physics	30	50	73	56	209

In order to demonstrate the performance of the proposed approach, the professional report on Ranked Journal List (RJL) [40] is adopted as a benchmark for comparison. The RJL provided by ERA 2010 involved a large group of scholars to rank a large number of academic journals. Despite that much debate surrounds the end result of RJL and other subjective forms of journal ranking, the ranking results provided by human experts have been frequently employed as benchmarks to compare journal ranking outcomes produced by automated mechanisms [13], [16]. RJL is also employed in this work, although it is not to serve as the gold standard for evaluating the performance of the proposed approach. Instead, it is used to demonstrate comparable ranking results with those provided by human experts, showing the potential similarity and difference between the result of the proposed data-driven method and that of peer-reviews. Indeed, RJL may involve biased human subjectivity which the present data-driven approach is to help avoid.

Each journal in RJL has a rank in the (ordered) domain $Ranks$ = {C, B, A, A*}, where rank A* indicates the top category of journals in a certain research area, and the significance and popularity of journals are decreasing from A* to C. When examining the selected indictor scores from JCR and the ranked result from RJL, only those journals that are both indexed by JCR and ranked in RJL are considered as valid experimental data. This is necessary to ensure that each journal used in the experiments has an external reference rank, to entail fair comparison. If a journal is missed from either RJL or JCR, then it is removed from the experimental data. A summary of the resultant datasets is shown in Table I. Each of these datasets contains over two hundred journals.

TABLE II
r_s COEFFICIENTS BETWEEN INDIVIDUAL INDICATORS AND RJL RANKS

ID	TC	IF	5-IF	II	Ei	AI
D1	0.6665	0.7962	**0.8152**	0.7557	0.7019	0.8125
D2	0.4914	0.4603	0.5023	0.3188	0.4130	**0.5480**
D3	0.6463	0.6153	0.6413	0.6045	0.6634	**0.7185**
D4	0.5923	0.5610	0.5884	0.5262	0.6427	**0.7287**
D5	**0.5401**	0.4961	0.5010	0.5083	0.5368	0.5375
D6	0.4501	0.6659	0.7299	0.5586	0.5095	**0.7614**
Ave.	0.5645	0.5991	0.6297	0.5454	0.5779	**0.6844**

B. Prior Analysis on Indicator Correlations

In statistics, Spearman's rank correlation coefficient r_s is a nonparametric measure of statistical dependence between two given variables [41]. It assesses how well the relationship between the two variables can be described using a monotonic function. If there are no identical data points, a perfect Spearman correlation of +1 or -1 occurs when each of the variables is a perfect monotonic function of the other. The sign of r_s indicates the direction of association between one variable, say x (calling it the independent variable) and the other, say y (the dependent variable). If y tends to increase when x increases, r_s is positive, and if y tends to decrease when x increases, r_s is negative. $r_s = 0$ indicates that there is no tendency for y to either increase or decrease as x increases. The r_s between scores of each individual indicator and the RJL ranks are listed in Table II. It shows that these indicators have a positive r_s value with respect to the RJL scores. This indicates that generally, if the scores of a journal on these indicators tend to increase, then their ranks in RJL increase also. However, for each indicator, its correlation levels to RJL are different from dataset to dateset. From their average performance on these datasets it can be seen that AI is the most correlated indicator to the rank of RJL, while II is the least relevant indictor. IF and 5-IF, which are commonly used in many real-world quality assessment scenarios, are more highly relevant to the results of RJL, as compared to TC and Ei.

To examine the results further, the correlations between individual indicators are computed as listed in Tables III-VIII. It can be seen from these tables that amongst the six indicators provided by JCR 2010, IF and 5-IF have the highest r_s coefficient and AI is also highly r_s-correlated to 5-IF, while TC is highly r_s-correlated to Ei. Both indicators of Total Cites (TC) and Eigenfactor (Ei) are biased towards journals which publish more papers, since they are not normalised with regard to the number of papers published within a certain period. The calculation of TC does not exclude self-citations while that of Ei does. However, it can be seen from the r_s coefficient between TC and Ei (and also that between 5-IF and AI), self-citations do not lead to any significant difference in ranking journals on these datasets. It can also be seen from these tables that TC and Ei form one neighbourhood while AI, 5-IF and IF form another, if trying to cluster these indicators. Finally, it is worth noticing that in general, the indicator II forms a neighbourhood of its own, though regarding the Mathematics dataset, it is closer to 5-IF and IF than AI.

TABLE III
r_s COEFFICIENTS BETWEEN INDICATORS – CHEMISTRY

	TC	IF	5-IF	II	Ei	AI
TC	1	0.6844	0.6697	0.6906	**0.9595**	0.6340
IF		1	**0.9805**	0.8900	0.7512	0.9565
5-IF			1	0.8834	0.7301	**0.9801**
II				1	0.7416	0.8679
Ei					1	0.7036
AI						1

TABLE IV
r_s COEFFICIENTS BETWEEN INDICATORS – COMPUTER SCIENCE

	TC	IF	5-IF	II	Ei	AI
TC	1	0.6105	0.6581	0.4786	**0.9162**	0.6429
IF		1	**0.9367**	0.6397	0.5625	0.7367
5-IF			1	0.6047	0.6128	**0.8378**
II				1	0.4560	0.5230
Ei					1	0.7002
AI						1

TABLE V
r_s COEFFICIENTS BETWEEN INDICATORS – MATERIAL SCIENCE

	TC	IF	5-IF	II	Ei	AI
TC	1	0.7232	0.7347	0.6939	**0.9672**	0.7074
IF		1	**0.9772**	0.8384	0.7611	0.9072
5-IF			1	0.8367	0.7702	**0.9356**
II				1	0.7168	0.8172
Ei					1	0.7614
AI						1

TABLE VI
r_s COEFFICIENTS BETWEEN INDICATORS – MATHEMATICS

	TC	IF	5-IF	II	Ei	AI
TC	1	0.7313	0.7324	0.6740	**0.9426**	0.5789
IF		1	**0.9710**	**0.8077**	0.7243	0.7355
5-IF			1	0.8027	0.7310	0.7770
II				1	0.6757	0.6596
Ei					1	0.6765
AI						1

TABLE VII
r_s COEFFICIENTS BETWEEN INDICATORS – MEDICINE

	TC	IF	5-IF	II	Ei	AI
TC	1	0.6919	0.6956	0.6808	**0.9483**	0.7034
IF		1	**0.9770**	0.8284	0.7694	0.9447
5-IF			1	0.8294	0.7748	**0.9766**
II				1	0.7296	0.8201
Ei					1	0.7902
AI						1

To support systematic comparison, the quality levels of the journals that are awarded with respect to each of the individual indicators are aggregated using five different operators, namely: DOWA, kNN-DOWA and OWA with

TABLE VIII
r_s Coefficients between Indicators – Physics

	TC	IF	5-IF	II	Ei	AI
TC	1	0.5701	0.5481	0.5499	**0.9424**	0.4287
IF		1	**0.9671**	0.8299	0.6414	0.8871
5-IF			1	0.8024	0.6262	**0.9348**
II				1	0.6164	0.7869
Ei					1	0.5348
AI						1

TABLE IX
Averaging Weight of Each Indicator in OWA with Andness Weighting Vector

ID	TC	IF	5-IF	II	Ei	AI
D1	0.0002	-0.0045	-0.0478	-0.0194	0.0349	0.0366
D2	0.0470	-0.0045	-0.0333	0.0043	0.0061	-0.0196
D3	-0.0152	-0.0073	-0.0081	-0.0119	0.0080	0.0345
D4	-0.0011	-0.0023	0.0010	-0.0009	0.0072	-0.0039
D5	0.0021	0.0174	-0.0362	-0.0162	0.0214	0.0115
D6	-0.0083	-0.0015	-0.0183	0.0156	0.0161	-0.0036
Ave.	0.0041	-0.0005	-0.0238	-0.0048	0.0156	0.0093

TABLE X
Averaging Weight of Each Indicator in DOWA

ID	TC	IF	5-IF	II	Ei	AI
D1	-0.0062	-0.0055	-0.0018	0.0096	-0.0090	0.0129
D2	-0.0098	0.0099	0.0140	-0.0173	-0.0063	0.0095
D3	-0.0097	0.0048	0.0003	0.0030	-0.0058	0.0074
D4	-0.0076	0.0078	0.0059	0.0029	-0.0036	-0.0054
D5	-0.0080	0.0053	0.0033	-0.0045	-0.0117	0.0156
D6	-0.0072	0.0159	0.0110	-0.0210	-0.0054	0.0067
Ave.	-0.0081	0.0064	0.0055	-0.0046	-0.0070	0.0078

TABLE XI
Averaging Weight of Each Indicator in 3NN-DOWA

ID	TC	IF	5-IF	II	Ei	AI
D1	-0.0040	-0.0015	0.0029	0.0075	-0.0063	0.0014
D2	-0.0046	0.0114	0.0162	-0.0312	-0.0011	0.0093
D3	-0.0172	0.0092	0.0062	0.0050	-0.0111	0.0079
D4	-0.0093	0.0120	0.0094	0.0060	-0.0050	-0.0131
D5	-0.0100	0.0119	0.0120	-0.0087	-0.0152	0.0100
D6	-0.0069	0.0174	0.0157	-0.0345	-0.0017	0.0100
Ave.	-0.0087	0.0101	0.0104	-0.0093	-0.0067	0.0043

W_{mean}, $W_{andness}$ and W_{orness}. Scores of each indicator are (separately) normalised to $[0,1]$ before clustering and aggregation. The weighting vectors in the OWA operators are not weight-dependent, thus a pre-definition of them is required. Instead of using the simple W_{max} and W_{min}, W_{orness} and $W_{andness}$ are employed (which are derived from the so-called linear stress functions [42]). In particular, $W_{orness} = (0.29, 0.24, 0.19, 0.14, 0.09, 0.05)$, $W_{andness} = (0.05, 0.09, 0.14, 0.19, 0.24, 0.29)$ and $W_{mean} = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$, given that there are six indicators to be aggregated in each of the experiments carried out. Note that $W_{andness}$ is directly implemented as the reverse of W_{orness} [43].

C. Comparative Analysis of Ranking Results

Both DOWA and kNN-DOWA use dependent weighting vectors, and the resulting weights represent the reliability of the corresponding arguments. In DOWA, the reliability is measured by the similarity of an argument to the average of all arguments, while in kNN-DOWA, the reliability is done by the similarity of an argument to its k nearest neighbours. Since there are six indicators to be aggregated, the k in kNN-DOWA is set to 3, indicating that the majority of all 5 neighbours are considered. Tables IX-XI show the average weights that are computed for each impact indicator in OWA with $W_{andness}$, DOWA and 3NN-DOWA, respectively. Every entry in these tables is subtracted by $1/6$ (the average weight of each indicator) from its real value, so that a positive number means that the indicator is more highly weighted than the average, and that a negative one means that it is weighted lower than the average.

It can be seen from Table IX that Ei, AI and TC have positive weights, while 5-IF, II and IF have negative weights when a conjunctive aggregation is run. These results indicate that on most journals, Ei, AI and TC tend to give lower scores as compared to the other three indicators. More specifically, Tables X and XI desmonstrate that when either DOWA or 3NN-DOWA is utilised, the indicators IF, 5-IF and AI lead to higher scores, showing that they are considered more "reliable" when used with these two aggregation operators. Note that individually, each of these three indicators also gains a relatively high r_s coefficient to the RJL result (see Table II).

The r_s coefficients between the aggregated scores and the RJL results are depicted as the dot-lines in Fig. 3. On five out of the six datasets, 3NN-DOWA achieves the best or second best r_s results across all the five aggregation operators. However, its performance on the Mathematics dataset is not so good as those obtained using other aggregation operators. A possible reason is that the most RJL-relevant indicators are Ei, AI and TC on the Mathematics dataset while 3NN-DOWA puts more weight on IF, 5-IF and II. Similar to 3NN-DOWA, OWA with $W_{andness}$ also shows good results on these datasets, which indicates that the ranks produced by RJL are more like a conjunctive outcome of the impact indicators as opposite to a disjunctive outcome of them.

The solid lines in Fig. 3 show the r_s coefficients between the journal ranks obtained by FAC and those by RJL. The number of base clusters on each impact indicator is consecutively set from 2 to 11 (to support a wide range of comparative examinations). Since the direct aggregation of pure scores can provide a detailed rank, to entail an unbiased comparison, Eqn. (12) is employed to produce a ranking of the journals based on the final fuzzy clusters returned by FAC. As the fuzzy c-means algorithm starts with a random initialisation, each point on the solid lines is the average of 30 independent runs. However, as the impact of the initialisation of fuzzy c-means to any one dimensional dataset is small, the standard deviation of the results is very tiny. Therefore, standard deviations are omitted

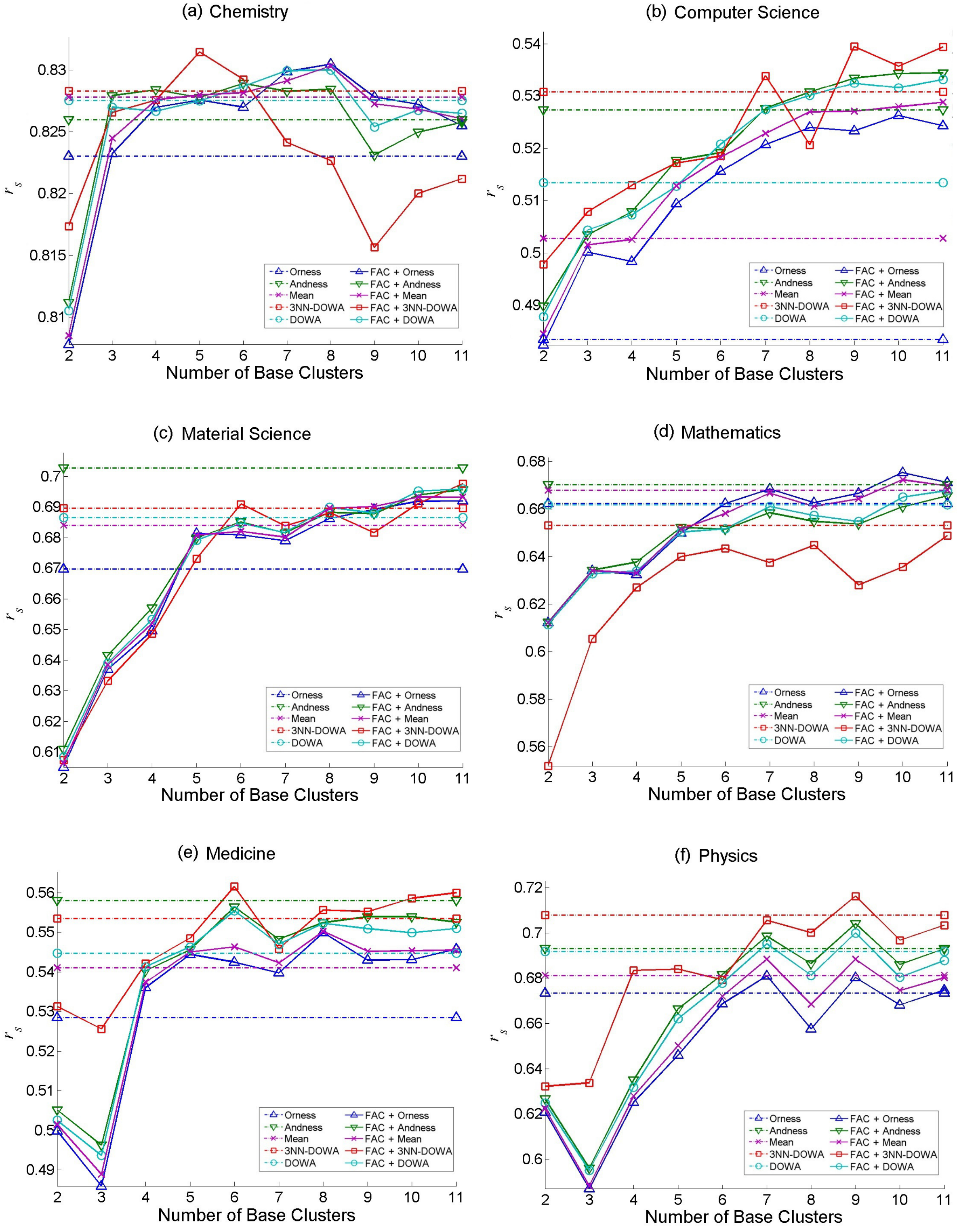

Fig. 3. Spearman's Correlation to RJL Results

from Fig. 3.

The first result to notice is that on five out of the six datasets, the solid lines can reach above the highest dotted lines. This indicates that using an appropriately selected number of base clusters, FAC can outperform the direct aggregation of individual indicator scores. These results also show that when FAC is employed, the highest r_s values on five out of the six datasets are achieved by the use of 3NN-DOWA. Overall, the results of 3NN-DOWA are better than those achievable using other aggregation operators on the following datasets: Computer Science, Medicine and Physics. Unfortunately, similar to the situation when 3NN-DOWA is applied to directly aggregate indicator scores, its performance on the Mathematics dataset is not so good as those obtained using other aggregation operators. Nevertheless, 3NN-DOWA generally achieves better results than DOWA both in direct aggregation and in FAC. According to Table II, AI has the highest r_s coefficient with RJL amongst all the six indicators. However, it can be seen from Table VI that the r_s coefficients between AI and other indicators are relatively low. In other words, AI cannot form a neighbourhood by itself to support its ranking, making the nearest neighbour approaches (both FAC+3NN-DOWA and 3NN-DOWA) fail to fit RJL.

In general, Fig. 3 does not show strictly monotonically decreasing or increasing patterns. Testing with more clusters may help further reveal the relationship between the number of base clusters and the ranking performance of the aggregated approach. However, there is a practical limitation on increasing the number of base clusters, especially in real application settings. Too many linguistic labels in use may hinder users' understanding of the differences between two adjacent labels. Therefore, the largest number of base clusters is set to 11 in the experimentation, which means that each base cluster will on average contain about 18 elements on a dataset with 200 journals.

Figure 3 shows that the results may vary when the number of the base clusters employed is increased. However, the differences between the highest and lowest Spearman's correlations to RJL are generally less than 0.1 in value across all datasets. That is, such variations are generally not significant. Examining the results more closely, it can be seen that the lowest values are always obtained when the number of base clusters is set to 2. When the defuzzication (defined by Eqn. (11)) is employed to refine the ranking results of FAC, the number of final clusters generated is the same as the number of base clusters for each indicator. When the weighted sum (defined by Eqn. (12)) is employed, the final ranks become stable if the number of base clusters is not too low.

To reflect this robustness property of the approach further, Table XII shows the averaged Spearman's correlation between the rank obtained by a given number of base clusters to that obtained by the use of a different number of base clusters (e.g., the robustness with #Cluster=2 is evaluated by the averaged Spearman's correlation between the rank of #Cluster=2 and each of those of #Cluster=3, 4, $\cdots$, 11). Clearly, the averaged Spearman's correlation values are in general very high. In particular, when the number of base clusters is above 4, the resultant ranks are very close to each other for all datasets, with the coefficients being greater than 0.96. If the number of base clusters is just 2, the ranking results deviate slightly from those obtained otherwise. However, practically speaking, it is not natural to employ only two base clusters in the first place. Thus, Table XII demonstrates that the number of base clusters does not affect the final ranks very much. In other words, the present approach is robust to the settings of this parameter. As such, plus the observation that each base cluster is labelled by a linguistic term in FAC, when given an application problem, the number of base clusters can be subjectively specified without adversely affecting the final ranking results significantly.

To facilitate further analysis of the experimental results on the proposed approach, Table XIII summarises the outcomes given in Fig. 3. Its first data column shows the best achieved r_s coefficients between FAC+3NN-DOWA and RJL, including an indication of where they have been achieved. The next column shows the same content except that the Total Cites indicator is removed from each dataset. The last column shows the mean and standard deviation of the r_s coefficients between the normal FAC+3NN-DOWA and that without TC. It can be seen from this table that the achieved results of the proposed method does not mirror the RJL results as much as the individual indicator scores (e.g., AI in Table II). This is expected because RJL is gained from subjective assessment and may contain biased human views that the proposed data-driven approach is to help avoid in the first place.

As an unsupervised approach, it is not surprising that the aggregated results deviate more from RJL than the most RJL-relevant indicators do. For example, as reflected in Table II, TC has the second lowest average r_s coefficient with RJL amongst the selected indicators. When TC is removed from the set of candidate indicators for aggregation, the r_s coefficient between the proposed method and RJL significantly decreases. A paired t-test is carried out between the results with or without TC, by changing the number of clusters from 2 to 11 across all datasets, and the t-test result is: 5.51×10^{-6}. This shows that removing TC will deviate the result of the proposed approach from RJL. However, comparing the r_s coefficients between the results with or without TC, it is clear that TC is not highly weighted in the proposed ranking. A possible reason for this is that TC only has one close neighbour (Ei), while AI, 5-IF and IF form the dominating neighbourhood (when the FAC+3NN-DOWA method is used).

Finally, it is worth noting that on datasets such as Computer Science and Medicine, none of the selected individual indicators has a high r_s coefficient to RJL. Therefore, the relatively low r_s of the aggregated results is not unexpected. This is partially because the RJL ranking is based on ratios of journals in a subcategory of each subject, such that an A* ranked journal in one subcategory could have a lower indicator score than an A ranked journal in another subcategory. After all, most of the journals are not significantly better or worse than others, although their ranks are more likely to be affected by the preference of the human assessors. In order to illustrate the eventual ranking results, as an example, Table XIV presents the top-10 ranked journals by the proposed method FAC+3NN-DOWA over the dataset D2 with the number of clusters set to 5. The ranking generally matches that as given by the human

TABLE XII
ROBUSTNESS OF FAC WITH DIFFERENT NUMBER OF BASE CLUSTERS

	Number of Base Clusters (#Cluster)									
ID	2	3	4	5	6	7	8	9	10	11
D1	0.9190	0.9588	0.9737	0.9783	0.9786	0.9782	0.9781	0.9747	0.9773	0.9745
D2	0.9378	0.9628	0.9737	0.9742	0.9752	0.9736	0.9753	0.9686	0.9665	0.9665
D3	0.9132	0.9506	0.9628	0.9740	0.9742	0.9740	0.9732	0.9744	0.9684	0.9712
D4	0.8989	0.9440	0.9667	0.9711	0.9706	0.9685	0.9711	0.9703	0.9682	0.9674
D5	0.9116	0.9354	0.9579	0.9672	0.9682	0.9701	0.9673	0.9648	0.9682	0.9632
D6	0.8929	0.9290	0.9576	0.9674	0.9634	0.9684	0.9695	0.9684	0.9670	0.9674
Avg.	0.9122	0.9468	0.9654	0.9720	0.9717	0.9721	0.9724	0.9702	0.9693	0.9684

TABLE XIII
COMPARISON OF r_s COEFFICIENTS WITH OR WITHOUT TC

	FAC vs. RJL (#Cluster)	FAC no-TC vs. RJL (#Cluster)	FAC vs. FAC no-TC
D1	0.8311 (5)	0.8280 (5)	0.9883 ± 0.0077
D2	0.5389 (9)	0.5245 (7)	0.9748 ± 0.0138
D3	0.6974 (11)	0.6780 (6)	0.9777 ± 0.0192
D4	0.6492 (11)	0.6383 (10)	0.9831 ± 0.0150
D5	0.5611 (6)	0.5460 (6)	0.9881 ± 0.0065
D6	0.7176 (9)	0.7138 (8)	0.9881 ± 0.0065

experts in ERA.

V. CONCLUSION AND FUTURE WORK

This paper has presented a fuzzy aggregation and clustering based method for academic journal ranking, focusing on the aggregated use of the impact indicators that appear in the Journal Citation Report provided by the Web of Science. The proposed method works by exploiting data-reliability based aggregation of fuzzy clusters that are generated from scores returned by individual impact indicators. It helps strengthen the interpretability of the assessment outcomes for academic journals, thanks to the use of quality level terms with inherent linguistic meaning. Experimental results on real-world journals from six subject areas have shown that the ranking results of the proposed method are generally consistent with those by RJL, which are produced by a large group of journal-ranking specialists. IImportantly, this is achieved without directly mirroring the rankings of RJL as the use of individual indicators may do, thereby helping to reduce the potential adverse impact of the bias introduced by subjective peer-reviews.

This promising research also opens up an avenue for significant further investigation. For instance, it would be useful to develop a method which would support aggregation of indicators involving different numbers of linguistic terms (i.e., different granularities of evaluation) [44]. Also, the present work is centred on the evaluation of journal impact indicators; it would be interesting to investigate whether the resultant techniques could be extended to coping with a broader range of problems, e.g., the assessment of the overall research quality of higher education institutions.

ACKNOWLEDGMENTS

The authors are very grateful to the Associate Editor and the reviewers for their constructive comments on the original manuscript which have helped improve this work significantly. The first and third authors are grateful to Aberystwyth University for providing them with PhD scholarships in support of this research. Part of this research has been supported by the Fundamental Research Funds for the Central Universities, China, under grant number: 2016MS120.

REFERENCES

[1] Research excellence framework. Higher Education Funding Council for England (HEFCE), the Scottish Funding Council (SFC), the Higher Education Funding Council for Wales (HEFCW) , the Department for Employment and Learning, Northern Ireland (DEL). [Online]. Available: http://www.ref.ac.uk/

[2] Excellence in research for australia (era). The Australian Research Council (ARC). [Online]. Available: http://www.arc.gov.au/era/

[3] I. R. Dobson, "Using data and experts to make the wrong decision," in *Using Data to Improve Higher Education*. Springer, 2014, pp. 229–242.

[4] J. Stegmann *et al.*, "How to evaluate journal impact factors," *Nature*, vol. 390, no. 6660, p. 550, 1997.

[5] E. Garfield, "The history and meaning of the journal impact factor," *JAMA: The Journal of the American Medical Association*, vol. 295, no. 1, pp. 90–93, 2006.

[6] C. Bergstrom, J. West, and M. Wiseman, "The eigenfactor metrics," *The Journal of Neuroscience*, vol. 28, no. 45, pp. 11 433–11 434, 2008.

[7] J. Jamali, M. Salehi-Marzijarani, and S. M. T. Ayatollahi, "Factors affecting journal quality indicator in scopus (scimago journal rank) in obstetrics and gynecology journals: a longitudinal study (1999-2013)," *Acta Informatica Medica*, vol. 22, no. 6, pp. 385–388, 2014.

[8] J. Stegmann and G. Grohmann, "Citation rates, knowledge export and international visibility of dermatology journals listed and not listed in thejournal citation reports," *Scientometrics*, vol. 50, no. 3, pp. 483–502, 2001.

[9] R. Rousseau, "Journal evaluation: technical and practical issues," *Library Trends*, vol. 50, no. 3, pp. 418–439, 2002.

[10] C. W. Holsapple, "A new map for knowledge dissemination channels," *Communications of the ACM*, vol. 52, no. 3, pp. 117–125, 2009.

[11] J. Bollen, H. Van de Sompel, J. A. Smith, and R. Luce, "Toward alternative metrics of journal impact: A comparison of download and citation data," *Information Processing & Management*, vol. 41, no. 6, pp. 1419–1440, 2005.

[12] B. Meyer, C. Choppy, J. Staunstrup, and J. van Leeuwen, "Viewpoint research evaluation for computer science," *Communications of the ACM*, vol. 52, no. 4, pp. 31–34, 2009.

TABLE XIV
TOP-10 JOURNALS IN D2 (COMPUTER SCIENCE)

Rank	Abbreviated Journal Title	ISSN	TC	IF	5-IF	II	Ei	AI	RJL
1	IEEE T PATTERN ANAL	0162-8828	25060	5.308	7.534	0.625	0.04969	2.802	A*
2	MIS QUART	0276-7783	7419	5.041	9.821	0.737	0.00926	2.760	A*
3	INT J COMPUT VISION	0920-5691	9898	5.151	6.986	0.808	0.02168	2.648	A
4	ACM COMPUT SURV	0360-0300	2888	8.000	10.91	0.867	0.00567	4.366	A*
5	IEEE T INFORM THEORY	0018-9448	28880	2.728	4.313	0.423	0.06987	1.691	A*
6	IEEE T MED IMAGING	0278-0062	11114	3.639	4.438	1.012	0.01877	1.258	A*
7	J CHEM INF MODEL	1549-9596	9556	3.822	3.722	0.756	0.01894	0.797	A
8	IEEE T IMAGE PROCESS	1057-7149	12774	2.918	4.205	0.333	0.03571	1.560	A*
9	J MACH LEARN RES	1532-4435	4766	2.974	4.967	0.456	0.02175	2.448	A
10	J ACM	0004-5411	6116	3.375	4.019	0.500	0.00706	2.347	A*

[13] G. Beliakov and S. James, "Citation-based journal ranks: the use of fuzzy measures," *Fuzzy Sets and Systems*, vol. 167, no. 1, pp. 101–119, 2011.

[14] S. Cooper and A. Poletti, "The new era of journal ranking," *Australian Universities Review*, vol. 53, no. 1, pp. 57–65, 2011.

[15] P. Su, C. Shang, and Q. Shen, "Link-based approach for bibliometric journal ranking," *Soft Computing*, vol. 17, no. 12, pp. 2399–2410, 2013.

[16] F. Bartolucci, V. Dardanoni, and F. Peracchi, "Ranking scientific journals via latent class models for polytomous item response data," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 178, no. 4, pp. 1025–1049, 2015.

[17] D. T. Anderson, J. C. Bezdek, M. Popescu, and J. M. Keller, "Comparing fuzzy, probabilistic, and possibilistic partitions," *Fuzzy Systems, IEEE Transactions on*, vol. 18, no. 5, pp. 906–918, 2010.

[18] R. Diao and Q. Shen, "Feature selection with harmony search," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 6, pp. 1509–1523, 2012.

[19] Q. Shen, R. Diao, and P. Su, "Feature selection ensemble," in *Alan Turing Centenary*, A. Voronkov, Ed., 2012, pp. 289–306.

[20] R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 18, no. 1, pp. 183–190, 1988.

[21] J. Malczewski, "Ordered weighted averaging with fuzzy quantifiers: Gis-based multicriteria evaluation for land-use suitability analysis," *International Journal of Applied Earth Observation and Geoinformation*, vol. 8, no. 4, pp. 270–277, 2006.

[22] C. Rinner and J. Malczewski, "Web-enabled spatial decision analysis using ordered weighted averaging (owa)," *Journal of Geographical Systems*, vol. 4, no. 4, pp. 385–403, 2002.

[23] J. J. Dujmović, "Properties of local andness/orness," in *Theoretical Advances and Applications of Fuzzy Logic and Soft Computing*. Springer, 2007, pp. 54–63.

[24] T. Boongoen and Q. Shen, "Clus-dowa: A new dependent owa operator," in *Fuzzy Systems (FUZZ-IEEE), 2008 IEEE International Conference on*. IEEE, 2008, pp. 1057–1063.

[25] Z. Xu, "Dependent owa operators," in *Modeling Decisions for Artificial Intelligence*. Springer, 2006, pp. 172–178.

[26] T. Boongoen and Q. Shen, "Nearest-neighbor guided evaluation of data reliability and its applications," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 6, pp. 1622–1633, 2010.

[27] G. Beliakov and S. James, "Using choquet integrals for knn approximation and classification," in *Fuzzy Systems (FUZZ-IEEE), 2008 IEEE International Conference on*. IEEE, 2008, pp. 1311–1317.

[28] R. Yager, "Families of owa operators," *Fuzzy Sets and Systems*, vol. 59, no. 2, pp. 125–148, 1993.

[29] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*. CRC Press, 2013.

[30] S. Lloyd, "Least squares quantization in pcm," *Information Theory, IEEE Transactions on*, vol. 28, no. 2, pp. 129–137, 1982.

[31] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2, pp. 191–203, 1984.

[32] K. Punera and J. Ghosh, "Consensus-based ensembles of soft clusterings," *Applied Artificial Intelligence*, vol. 22, no. 7-8, pp. 780–810, 2008.

[33] P. Su, C. Shang, and Q. Shen, "Link-based pairwise similarity matrix approach for fuzzy c-means clustering ensemble," in *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1538–1544.

[34] N. Otsu, "A threshold selection method from gray-level histograms," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 9, no. 1, pp. 62–64, 1979.

[35] N. Iam-On and T. Boongoen, "Comparative study of matrix refinement approaches for ensemble clustering," *Machine Learning*, vol. 98, no. 1-2, pp. 269–300, 2015.

[36] G. Beliakov, A. Pradera, and T. Calvo, *Aggregation functions: A guide for practitioners*. Springer, 2008.

[37] T. Calvo, G. Mayor, and R. Mesiar, *Aggregation operators: new trends and applications*. Springer, 2002, vol. 97.

[38] J. F. Kolen and T. Hutcheson, "Reducing the time complexity of the fuzzy c-means algorithm," *Fuzzy Systems, IEEE Transactions on*, vol. 10, no. 2, pp. 263–267, 2002.

[39] L. Leydesdorff, "Can scientific journals be classified in terms of aggregated journal-journal citation relations using the journal citation reports?" *Journal of the American Society for Information Science and Technology*, vol. 57, no. 5, pp. 601–613, 2006.

[40] Ranked journal list development. The Australian Research Council (ARC). [Online]. Available: http://www.arc.gov.au/era/journallistdev.htm/

[41] G. W. Corder and D. I. Foreman, *Nonparametric Statistics: A Step-by-step Approach*. John Wiley & Sons, 2014.

[42] R. Yager, "Using stress functions to obtain owa operators," *Fuzzy Systems, IEEE Transactions on*, vol. 15, no. 6, pp. 1122–1129, 2007.

[43] R. Yager, "Generalized owa aggregation operators," *Fuzzy Optimization and Decision Making*, vol. 3, no. 1, pp.93–107, 2004.

[44] Q. Shen and T. Boongoen, "Fuzzy orders-of-magnitude-based link analysis for qualitative alias detection," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 4, pp. 649–664, 2012.

Pan Su received the M.Sc. degree in computer science and technology from Hebei University, Hebei, China, and the Ph.D. degree in computer science from Aberystwyth University, Aberystwyth, U.K. in 2011 and 2015, respectively. He is a lecturer with the School of Control and Computer Engineering, North China Electric Power University, Baoding, China. His research interests include fuzzy systems, fuzzy aggregations, machine learning, and data mining.

Changjing Shang received a Ph.D. in computing and electrical engineering from Heriot-Watt University, UK. She is a University Research Fellow with the Department of Computer Science, Institute of Mathematics, Physics and Computer Science at Aberystwyth University, UK. Prior to joining Aberystwyth, she worked for Heriot-Watt, Loughborough and Glasgow Universities. Her research interests include pattern recognition, data mining and analysis, space robotics, and image modelling and classification.

Tianhua Chen received the B.Sc. degree in software engineering from Fujian Normal University, Fuzhou, China, and the M.Sc. degree in artificial intelligence from Aberystwyth University, Aberystwyth, U.K. in 2012 and 2013, respectively. He is currently working towards the Ph.D. degree in computer science with the Department of Computer Science, Institute of Mathematics, Physics and Computer Science, Aberystwyth University, Aberystwyth, U.K. His research interests include fuzzy systems, fuzzy set theory, and pattern recognition.

Qiang Shen received the Ph.D. degree in computing and electrical engineering and the D.Sc. degree in computational intelligence. He holds the Established Chair in Computer Science and is the Director of the Institute of Mathematics, Physics and Computer Science, Aberystwyth University, Aberystwyth, U.K. His research interests include computational intelligence, reasoning under uncertainty, pattern recognition, data mining, and real-world applications of such techniques for intelligent decision support (e.g., crime detection, systems monitoring, medical diagnosis, and quality assessment). Of direct relevance to the work presented in this paper, Professor Shen was a panel member for the UK REF 2014 on Computer Science and Informatics. He has authored two research monographs and more than 350 peer-reviewed papers in these areas, including an Outstanding Transactions Paper Award from this journal.