International Journal of Engineering & Technology, 7 (2.7) (2018) 28-32

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

A critical study and analysis of journal metric "CiteScore", cluster and regression analysis

K. Varada Rajkumar[1]*, Yesubabu Adimulam[2], K. Subrahmanyam[3]

[1]KLEF, India
[2]Sir C R Reddy College of Engineering, India
[3]KLEF, India
**Corresponding author E-mail: varadarajkumar18@gmail.com*

Abstract

In previous days the quality of journal is measured based on many metrics such as impact factor; SNIP (Source Normalized Impact Per paper),SJR(*SCImago Journal Rank*) and IPP(Impact Per Publication). It is very hard to find that the research papers to be published in which journal? CiteScore is a better way of measuring the citation impact of sources such as journals. CiteScore is a metrics product for journal from elsevier, using the citation data from the scopus database to rank journals. CiteScore metrics is a comprehensive current and free metrics for source titles in scopus. Apart from Impact factor, CiteScore is becoming increasingly important in the context of evaluating metrics for all journals. CiteScore metrics are available for 37956 titles in scopus. It is not limited to journals as also conference proceedings, trade, publications and book series. The metrics are available 6 years period from 2011 to 2016. For a subset of CiteScore dataset clustering and regression algorithms can be implemented to study the data points that lie equally distant from one another.

Keywords: *Jounral Metrics; CiteScore; Clustering Algorithms; Fuzzy Clustering; Regression.*

1. Introduction

CiteScore is the ratio of citation of document in this year to the total citations in the previous three years. The importance of CiteScore is increasing in the context of evaluating the journals from the past six years. CiteScore is comprehensive transparent and current.

Cluster analysis can be described as a statistical tool which is used to differentiate data into groups. There are plenty of clustering algorithms which are used to categorize the data into groups. Clustering can be considered to be one of most significant solo learning problem; like the other problem of this type, it finds the structure for a group of unlabelled data [1]. A straight forward definition of clustering would be "the process of organizing objects into groups whose members are similar in some way". Hence A cluster is a group of objects which are similar and are "dissimilar" to the objects which belong to other clusters. It is easy for us to identify 4 clusters into which the data can be distributed, if the criterion is the distance between the two objects that is the geometrical distance between two points. If two objects are included in same cluster, which means they are very close according to the given geometrical distance. This process of dividing the objects into clusters based on distance criterion is called distance-based clustering[2] [3].

Regression analysis is a statistical modeling tool used to estimate the relation among the predictor and response variables. One of the most important characteristics of regression models is their predictive power. The accurate prediction of target property which are not used for model development is another significant characteristic of regression analysis. This is attained by rational partition of an experimental SAR dataset into the training and test set, which are used for model development and validation, respectively [4]. In a specified way regression analysis describes how the value of response variable is changed for different values of predictor variables. For every instance regression function is calculated which is a function of predictor variable. Regression analysis is mainly used in the field of machine learning , mainly in prediction and forecasting. Many techniques have been developed for regression analysis. Most recognized methods are linear regression and ordinary least squares regression.

2. Literature Review

2.1. CiteScore

In determining the quality of academic journals and scholarly publishing the journal impact factor (IF), owned by Thomson Reuters has been a leading metric. Impact factor is not used for choosing a journal in which to publish or as a measure for finding journal quality, it is designed to help the librarians to make decisions about their journal collections. The Impact factor dominated scholarly publishing for the part of six decades. The journal's Impact factor is a method of calculating the number of times the articles in a journal indexed in the Web of Science database are cited within 2 years period in other journals of the same database[5]. The comparison between disciplines should not be done using impact factor. The practice of the Citations are based on its subject, with this the result will be a high impact factor for a subject may look tremendously low when compared with another subject. To undergo this disadvantage Elsevier B.V. , a direct competing metric is launched to the Impact Factor, CiteScore (CS)[6] On December 8, 2016. CiteScore calculates the total number of citations of all documents in first year to total documents publicized in the past three years on a particular label. It is a

very vigorous and specific sign of a journal's impact. It can be briefly defined as the mean of citations for a single document that a title sustains in three years. If is depends on the Web of Science database where the CiteScore is based on the journals in Scopus. The Scopus contains twice as many journals as Web of Science database. CiteScore is calculated monthly and is based on citations from about 22,000 sources. CiteScore metrics are a collection of following complementary indicators. Fig. 1. gives the values for CiteScore metrics for some journals in 2016.

1. CiteScore tracker
2. CiteScore Percentile
3. CiteScore Quartiles
4. CiteScore Rank
5. Citation Count
6. Document Count
7. Percentage Cited

CiteScore tracker

CiteScore tracker shows the building up of current year's CiteScore each month.

CiteScore Percentile

The journals are compared in subject areas using the CiteScore percentile which normalizes the CiteScore within the discipline. The CiteScore percentile runs from a highest rank of 100 down to lowest rank of 1 .

CiteScore Quartiles

Because of occupying the same position within their subject categories,band of serial titles are grouped together to form quartiles.

CiteScore Rank

CiteScore rank provides the absolute standing of the title of journal in its particular field

Citation Count

The ratio of sum of citations encountered in one year to the documents published in 3 preceding years can be described as Citation Count.

Document Count

It is defined as the sum of documents published with serial title in last three years to the year the metric is calculated.

Percentage Cited

The proportion of the documents present in the denominator of the CiteScore calculation that have received at least 1 citation in the numerator.

	Title	CiteScore	CiteScore Percentile	CiteScore Rank	Citations 2016	Documents 2013-15	% Cited	SNIP	SJR
1	Foundations and Trends in Signal Processing *Signal Processing*	23.00	99%	1/89	92	4	75%	14.072	2.902
2	IEEE Transactions on Pattern Analysis and Machine Intelligence *Software*	13.29	99%	1/367	8,189	616	90%	6.317	6.298
3	IEEE Transactions on Pattern Analysis and Machine Intelligence *Computer Vision and Pattern Recognition*	13.29	99%	1/66	8,189	616	90%	6.317	6.298
4	IEEE Transactions on Pattern Analysis and Machine Intelligence *Computational Theory and Mathematics*	13.29	99%	1/97	8,189	616	90%	6.317	6.298
5	IEEE Transactions on Pattern Analysis and Machine Intelligence *Artificial Intelligence*	13.29	99%	1/152	8,189	616	90%	6.317	6.298
6	Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition *Software*	12.72	99%	2/367	20,576	1,617	92%	4.551	7.342
7	Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition *Computer Vision and Pattern Recognition*	12.72	97%	2/66	20,576	1,617	92%	4.551	7.342
8	IEEE Transactions on Evolutionary Computation *Software*	11.83	99%	3/367	2,023	171	96%	5.404	3.544

Fig. 1. The values for CiteScore metrics for some computer science journals in 2016. Source: Scopus**.**

In the last five years the abrupt raise in the relative and absolute terms in the citations of the document in the last five years, according to recordings in Scopus is represented in Fig. 2. The CiteScore giving the place of 88th in percentile it has been in-

creased to 2.40 in 2016 in the Finance Subject, and it ranked 25 in all 216 journals in that subject (Table 1). As an inherent measure, this shows that the rank of Venture Capital is increasing rapidly in its subject Finance[7].

Table 1. Ranks of Venture Capital using CiteScore

Year	CiteScore	Percentile	Journal rank in finance category
2012	0.79	57	79/187
2013	0.75	50	94/189
2014	0.75	50	92/194
2015	1.87	85	30/206
2016	2.4	88	25/216

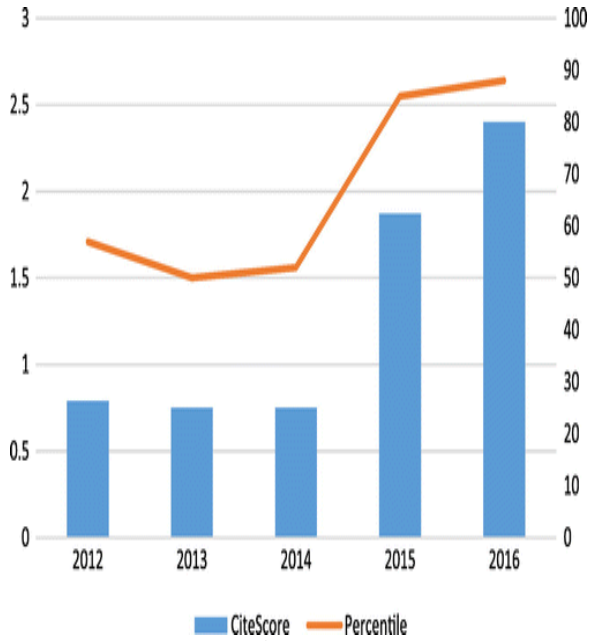

Fig. 2. CiteScore Rankings for Venture Capital. Source: Scopus.

Features of CiteScore

Comprehensive

CiteScore is defined as the average citations that a title receives in the past three year period per document. It is very easy to replicate.

Transparent

The method of calculating the CiteScore is straight forward with no hidden details or secret algorithms. It gives more accurate and robust indication of a journals impact.

Current

CiteScore ,showing the citations for full calendar year ,is calculated on annual basis. CiteScore Tracker provides the current view of journal's performance during the course of year[8].

Calculation of CireScore

CiteScore counts the citations of documents received in the current year followed by the counting the citations in the previous three years. Then the ratio of these two the value of CiteScore. let A be the value of citations received by the documents published in 2016 and B is the value of citations in 2013,2014,2015. The CiteScore value will be A/B. Fig. 3. Show the Calculation of CiteScore. Fig. 4. Show the Calculation of CiteScore of "IEEE Transactions on Pattern Analysis and Machine Intelligence" Journal in the year 2016.

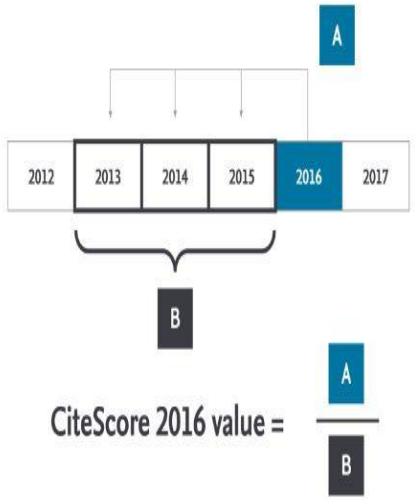

Fig. 3. Calculation of CireScore

Fig. 4. CiteScore of IEEE Transactions on Pattern Analysis and Machine Intelligence" Journal in the year 2016

2.1. Clustering Analysis

The process of maximizing the intraclass similarity and minimizing the interclass similarity. Clusters are formed so that the object belonging to the same cluster which contain similar data and the objects with dissimilarity are placed in different clusters.

Classification of clustering

Clustering is classified into following subgroups.

1. Hard clustering
2. Soft Clustering
3. Hierarchical clustering
4. Partition clustering
5. Exclusive Clustering
6. Overlapping Clustering
7. Fuzzy Clustering
8. Complete Clustering

Hard clustering

The process of clustering in which every data point is either belong to a cluster totally or not is called hard clustering. In hard clustering clusters do not overlap.

Soft Clustering

The process in which the data points are not placed in separate clusters instead, the probability of the cluster is assigned is called soft clustering.

Hierarchical clustering

Hierarchical clustering exists as a cluster in a bigger cluster to form a tree. As a result, the hierarchical clustering is also known as nested clustering.

Partition clustering

The process of dividing the set of data objects such that each object should consists of exactly one subset. In partition clustering the clusters will not overlap.

Exclusive Clustering

Exclusive clustering deals with the assignment of each value to only one cluster.

Overlapping Clustering

Overlapping clustering is used to shrine up the aspect that an object can concurrently belong to more than one group.

Fuzzy clustering

In fuzzy clustering, the concept of membership weight comes into existence. Here every object will be a part of every cluster. The membership weight that goes between 0:if it utterly doesn't belong to cluster and 1:if it utterly belongs to the cluster[9**].**

Complete clustering

The task of performing the hierarchical clustering using a set of dissimilarities on 'n' objects that are being clustered is called Complete clustering. They tend to find dense clusters of an approximately equal diameter.

Clustering algorithms

As the task of performing clustering is intuitive, which means of achieving the goals are plenty. We have known more than 100 clustering algorithms until now. Some of the algorithms are K-means, Hierarchical clustering, Fuzzy C-means and Mixture of Gaussians.

Classification of Clustering Algorithms

In detail there are 4 kinds of clustering algorithms

Connectivity Models

The name itself indicate that in these models the data points which are closer to each other will have high similarity and can be placed in same cluster and the data points than the data points farther from each other. The connectivity models basically be done in two ways. The first way starts with categorizing all the data points into different clusters and later aggregating them as the interval between them reduces. While in next way, all the data points are categorized as one cluster and then divided as the interval between them raises. Selection of distance function is intuitive. Connectivity model is lacked behind in the concept of scalability for handling big datasets and are very easy to interpret. hierarchical clustering algorithm and its variants are the best examples of Connectivity models.

Centroid models

Centroid models are iterative clustering algorithms. The similarity of the data points are obtained by nearness of a data point from the centroid of the cluster. The popular algorithm, K-Means clustering algorithm is the best example of Centroid model. the number of clusters needed at last should be defined before hand, which makes it significant to have previous knowledge on the dataset. To know the local ,maxima of the centroid models, they run iteratively.

Distribution models

Distributed clustering models depends on how feasible are the data points present in the cluster, which is a part of the similar distribution. Normal and gaussian algorithms fall under this model. Distributed models often suffer from over fitting. The main algorithm of multivariate normal distributions, expectation-maximization algorithm is example of Distributed model.

2.2 Density Models

Density models explore the areas of diverse density of data points in the data space . Density models separates a variety of contrast density areas and allot the data points in those regions in the similar cluster. DBSCAN and OPTICS are the popular examples of density models.

Applications of Clustering

Clustering has several applications in various domains. They are

- Recommendation engines
- Market segmentation
- Analysis of Social networks
- Grouping of search results
- Medical imaging
- Image segmentation
- Outlier detection

2.3. Regression analysis

The predictive modeling technique which look over the relationship between a one or more response variables and predictor

variables is called Regression analysis[10]. Regression analysis is one of the best tools used to model and analyze the data. In the below Figure we sketch a curve or a line for the data points(Fig. 5.) in a way that the figure represents the contrast between the distances of data points from the curve or line is reduced.

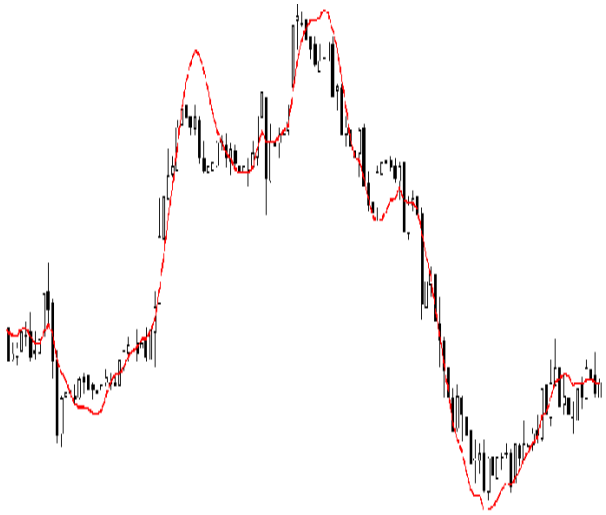

Fig. 5. Curve or lines for the data points.

Importance of Regression analysis

Regression analysis is the remarkable relationship between predictor and response variables . It also specifies the supremacy of impact of multiple predictor variables on response variables. It compares the effects of variables measured on dependent variables.

Types of Regression analysis

Linear regression:

One of the most extensively used modeling techniques is Linear Regression. Linear Regression set a relation in-between the response variables and one or more predictor variables through a straight line called Regression line.

Logistic Regression:

The use of Logistic Regression is to calculate the probability of either favorable outcomes or unfavorable outcomes. The value of that will be from 0 to 1. The value can be deducted using the formulae

Odds = k/(1-k)

where k = probability of favorable outcomes

and 1-k = probability of unfavorable outcomes.

ln(odds)=ln(k/(1-k))

logit(k)=ln(k/(1-k)) = a0+a1x1+a2x2+.....akxk

Polynomial Regression

If the power of predictor variables are greater than 1 then the regression equation is called polynomial regression equation. the equation is represented as
y=a+b*x^2

Step wise regression

While handling the numerous predictor variables stepwise regression is used. An automatic process which involves no humans is used for the selection of independent in this stepwise regression.

Ridge Regression

When the data has multi co linearity (predictor variables are highly correlated) Ridge Regression is used. In ridge Regression an error term is used. The equation used for ridge regression is
Y=a+b*x+e

Lasso Regression

Lasso (least absolute shrinkage and selection operator) regression like ridge regression deals with the absolute size of the Regression Coefficients. The only dissimilarity in lasso and ridge regression is the usage of absolute values by lasso.

ElasticNet Regression

Elastic net regression is defined as the hybrid of least absolute shrinkage and selection operator regression and ridge regression. Elasticnet Regression suffers with double shrinkage. When it comes to the point that multiple features are to be correlated then Elastic net is very useful.

3. Conclusion

In this paper, a brief review is done for the importance of CiteScore, calculating CiteScore value, different clustering algorithms and regression algorithms. CiteScore is becoming increasingly important in the context of evaluating metrics for all journals. Taking CiteScore dataset and applying clustering and algorithms they are seemed to be lie equally distant from one another. By doing the regression analysis on CiteScore dataset may be find how dependent variable depends on independent variables.

References

[1] Brian Everitt, “Cluster Analysis,” 2nd Edition, chapter 3. Halsted Press, 1980.

[2] Kirsten M, Wrobel S., “Relational distance-based clustering” Inductive Logic Programming”, pp. 261-70,1998.

[3] Liu J, Mohammed J, Carter J, Ranka S, Kahveci T, Baudis M. “Distance-based clustering of CGH data,” Bioinformatics, Vol. 22, Iss. 16, pp. 1971-1978, 2006.

[4] Fox J. “Applied regression analysis, linear models, and related methods”, Sage Publications, Inc; 1997.

[5] SA Azer A Holen, I Wilson, and N Skokauskas , “Impact factor of medical education journals and recently developed indices: Can any of them support academic promotion criteria?,” Journal of Postgraduate Medicine, Jan-Mar; 62(1), pp. 32–39, 2016.

[6] Jaime A.Texixeria da Silva, Aamir Raoof Memon, “CiteScore: A cite for sore eyes, or a valuable, transparent metric?,” Scientometrics, Vol 111, Iss. 1, pp. 553-556, 2017.

[7] Richard T. Harrison, “Signalling journal impact and prestige: Venture Capital: An International Journal of Entrepreneurial Finance,” Venture Capital, An International Journal of Entrepreneurial Finance, Vol 19, Iss 4, pp. 257-262, 2017.

[8] Hans Zijlstra , Rachel McCullough, “CiteScore: a new metric to help you track journal performance and make decisions,” ELSEVIER, 2016.

[9] P. Alam, D. Booth, K. Lee, T.Thordarson, “The use of fuzzy clustering algorithm and self-organizing neural networks for identifying potentially failing banks: an experimental study,” Expert Systems with Applications, Vol. 18, Iss. 3, pp. 185-199, 2000.

[10] Yixin Chen, Guozhu Dong, Jiawei Han, Benjamin W.Wah and Jianyoung Wang, “Multi-Dimensional Regression Analysis of Time-Series Data Streams,” Proceeding of the 28th International Conference on Very Large Databases, pp.323 – 334, 2002.