DOI: 10.15439/2023F2798
ISSN 2300-5963 ACSIS, Vol. 36

Evaluation of selected Cardinality Pattern functions and linguistic variables applied to authors dominant discipline classification

Lukasz Szymula
0000-0001-8714-096X
(1) Faculty of Mathematics and Computer Science
(2) Center for Public Policy Studies, Adam Mickiewicz University in Poznan, Poland
(3) Department of Computer Science, University of Colorado Boulder, United States
Email: lukasz.szymula@amu.edu.pl

Krzysztof Dyczkowski
0000-0002-2897-3176
Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznan, Poland
Email: krzysztof.dyczkowski@amu.edu.pl

***Abstract*—The ongoing study aimed to investigate the impact of utilizing intelligent counting algorithms to determine the dominant discipline of authors. This paper addresses the issue of ambiguously assigning disciplines to authors, which has become a prevalent problem. The methodology section outlines the approach employed in this study, including the utilization of intelligent counting, cardinality pattern functions, and evaluation metrics. In the results section, we present the findings of the study, demonstrating that by employing specific Cardinality pattern functions and linguistic variables, we were able to achieve a return that surpassed the number of disciplines unambiguously determined for authors by up to 30%, surpassing the results obtained using well-known methods.**

***Index Terms*—intelligent counting, cardinality pattern functions, science of science, determining disciplines**

I. Introduction

IN THE field of Scientometrics, research is conducted at multiple levels, focusing on various units of analysis such as publications or researchers. These levels encompass diverse groups, including countries, disciplines, gender, age groups, sources, and research metrics. To ensure the selection of appropriate observation sets, it is crucial to assign unambiguous values to each observation. Ongoing research conducted at the Center for Public Policy Studies, Adam Mickiewicz University, indicates the need to exclude certain observations from the study due to the absence of specific values. Additionally, the challenge arises from the inability to assign a single, definitive value to each observation."

While for scientific databases in the case of some attributes there may be problems in indicating a specific value (complete impossibility to determine the value or appearance of outlier observations that underestimate the results), in the case of other dimensions alternative ways of determining the value can be considered. In Scientometrics, there are metrics for ranking publications and authors. They take values such as number of citations, number of publications, journal percentile, journal indices, author indices, and others at different depth level [7]. Scientific metrics allow ranking selected observations for the purpose of conducting further evaluation, and the prestige of the studied entity is determined by their placement. There are many metrics (CiteScore, FWCI, Percentile, H-Index, Collaboration metrics, etc.) and university rankings (Academic Ranking of World Universities, CWTS Leiden Ranking, RUR World University Rankings, etc.) [8], [13], [23], where terms like high, highest, most popular, medium, low, lowest, worst are used. These are imprecise terms, which are considered and modeled by the field of artificial intelligence: fuzzy set theory and linguistic variables proposed by L.A Zadeh [27].

II. Problem specification

For authors discipline determination Abramo, Aksnes, and D'Angelo, who defined the Web of Science subject category for each Italian and Norwegian professor in their sample [1] and Kwiek and Roszka [18] and Boekhout, van der Weijden and Waltman [5] who defined Scopus ASJC discipline for Polish scientists and global scientists, respectively, used an approach where they determined the modal value based on journals ASJC classification code. This method for one of the largest scientific database (Scopus), allowed 24,938,113 of the 36,010,088 (around 69%) authors to be classified into one dominant discipline thus leaving the remaining 11,071,975 (around 31%) with more than one discipline assigned (Tab. 1.). These 31 percent of observations are tend to be excluded from the samples. As an alternative to use of disciplines assigned to scientific journals and modal value assignment,

This work is the result of research project No. 2019/35/0/HS6/02591 funded by the National Science Center Poland and supervised by Professor Marek Kwiek.

techniques based on machine learning taking into account abstracts and keywords of publications have been used. Daradkeh M, Abualigah L, Atalla S, Mansoor W. have done paper field classification for Scopus, ProQuest, and EBSCOhost datasets using convolutional neural networks with titles, abstracts, keywords for papers and journal titles as a features [9]. Sood, S.K., Kumar, N. & Saini, M. have used the set of 7 thousand papers from Scopus database and clustering techniques based on keywords and VOSViewer environment [21]. Meen C. K., Seojin N., Fei W., Yongjun Z. performed graph analysis and text mining for keywords from 10 thousand articles from Web of Science database [19].

TABLE I.
DISTRIBUTION OF THE NUMBER OF SCIENTISTS BY NUMBER OF ASSIGNED DISCIPLINES

Number of disciplines	Number of authors
1	24,938,113
2	7,655,307
3	2,538,076
4	615,292
5	198,175
6	48,216
7	11,773
8	3,528
9	1,382
10	178
11	36
12	11
13	1
Total	36,010,088

With the occurrence of imprecise terms in the area of Scientometrics, a variety of possibilities arise to implement this algorithm. One way to measure sample set cardinality is to consider intelligent counting (Sigma f-Count) with various cardinality pattern functions proposed by Wygralak M. [26] and Dyczkowski K. [11]. Cases where author have more than one dominant discipline create space to introduce the fuzzy logic conceptual apparatus in order to increase the number of unambiguously identified authors.

The purpose of the ongoing study was to investigate the impact of intelligent counting usage in algorithm to determine authors dominant discipline. The study compares the sets of dominant author disciplines implemented using the crisp set approach and using fuzzy sets approach. The research questions are as follows:

1. How does the use of intelligent counting effect the number of uniquely classified authors?
2. Which linguistic variables, terms and cardinality pattern functions are meaningful in acquiring more unambiguously classified authors?
3. Does the result from using intelligent counting assign the same classes as classical approach?
4. Are there cardinality pattern functions that in any case return less observations than the approach using crisp sets.

III. METHODS

The bibliometric database Scopus from the ICSR Lab platform has been used for the study. Access to the database was granted through a collaboration between AMU's Center for Public Policy Studies and the International Center for the Study of Research (ICSR Lab), Elsevier, established in November 2020. The ICSR Lab allows access to the Scopus bibliometric database via the Databricks platform and retrieves results in aggregated form. Computations in the Databricks platform were based using cluster in standard mode with Databricks Runtime version 11.2 ML, Apache Spark technology v3.3.0, Scala v2.12, and instance i3.2xlarge with 61 GB Memory, 8 Cores, 1-6 workers for worker type and instance c4.2xlarge with 15 GB Memory, 4 Cores for Driver type. The execution time for all scripts took approximately 2 hours. Our sample included a set of authors and their dominant disciplines determined using an approach commonly used in Scientometrics (crisp sets, referred to by us as the base approach) and applying methods known from intelligent counting (selected cardinality pattern functions). The results were then evaluated following selected evaluation metrics.

A. Method for determining author's dominant discipline

The rule for determining author's dominant discipline was based on using a set of publications from the ICSR Lab platform. Each publication had its own unique identifier, a list of authors, a list of disciplines assigned to the journal from which the publication had come, and the variables Citation, FWCI 4y, FWCI 5y, FWCI NoWindow, Team size and Percentile. The Citation variable represented the total number of citations of the publication, the FWCI variables represented citation indices (gained up to 4, 5 years after the release date of the publication or without time limitation) normalized to the scientific discipline. Team size represented the number of authors in the publication, and the Percentile variable measured the percentile value of the CiteScore metric from the journal assigned to the publication. To determine an author's dominant discipline in the base approach for each author, the number of publications for each discipline that author had been counted. In the case of intelligent counting, each record was assigned to an appropriate membership degree based on the discipline, linguistics variable and term, and then in each discipline that the author subserved, the membership degrees were summed by relying on the appropriate cardinality pattern function and Sigma f-Count function. Then, for each author, only those disciplines were selected for which the number of publications (and Sigma *f*-Count score, respectively) were the highest. Multidisciplinary was excluded from the collection due to the fact that it is not a scientific discipline. The result was a set containing author identifier and his discipline. For the presented approach, the number of disciplines for each observation was greater than or equal to one.

B. Cardinality pattern functions and Sigma *f*-Count

For the purposes of this study to calculate the number of publications for each authors discipline we have used the sigma *f*-Count cardinality of a fuzzy set defined as:

$$\forall A \in FFS : sc_f(A) = \sum_{x \in supp(A)} f(A(x)),$$

where FFS is Family of all Fuzzy Sets, f is a cardinality pattern function, sc is scalar cardinality and $A(x)$ is interpreted as degree of membership of x to a fuzzy set A [26], [11]. As the cardinality pattern functions we decided to select four functions from the two patterns: counting by thresholding and counting by thresholding and joining.

1. $f_{1,t,p}$, where $t \in [0, 1]$ and $p \geq 0$. Called as counting by thresholding and joining by Wygralak [26]

$$f_{1,t,p}(x) = \begin{cases} x^p, & a \geq t, \\ 0, & otherwise. \end{cases}$$

2. $f_{2,t,p}$, where $t \in [0, 1]$ and $p \geq 0$. Called as counting by thresholding by Wygralak [26]

$$f_{2,t,p}(x) = \begin{cases} 1, & a \geq t, \\ x^p, & otherwise. \end{cases}$$

For cardinality patterns above we decided to use these in two combinations by p and 5 combinations by t which we named $f_{3,t}$ for $f_{1,t,p}$ with $p = 1$; $f_{4,t}$ for $f_{1,t,p}$ with $p = 2$; $f_{5,t}$ for $f_{2,t,p}$ with $p = 1$ and $f_{6,t}$ for $f_{2,t,p}$ with $p = 2$, where $t \in \{0, 0.2, 0.4, 0.6, 0.8\}$ giving 20 functions for each term of each linguistic variable (that is 360 calculations in total; 6 linguistic variables, 3 terms, 5 thresholds, 4 cardinality pattern functions) (Fig. 1.) [26], [11].

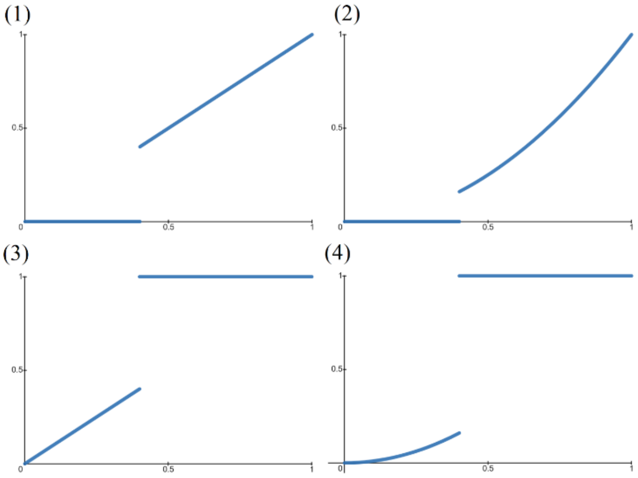


Figure 1. (1) Cardinality pattern $f_{1,t,p}$ for $t = 0.4$, $p = 1$ (as $f_{3,t}$), (2) Cardinality pattern $f_{4,t}$ for $t = 0.4$, $p = 2$ (as $f_{4,t}$), (3) Cardinality pattern f_2 for $t = 0.4$, $p = 1$ (as $f_{5,t}$), (4) Cardinality pattern f_2 for $t = 0.4$, $p = 2$ (as $f_{6,t}$)

C. Membership function modeling

Six linguistic variables (Citation, FWCI 4y, FWCI 5y, FWCI NoWindow, Percentile and Team size) with three terms (Low, Medium, High) in a universe covering the interval [maximum, minimum] for each linguistic variable in the Scopus database have been proposed for the determination of membership functions. The determination of the membership function was based on the same rule for each linguistic variable. In the case of term "high", the membership function was given a value 1 for the top 10 percentile of the variable's value. For the remaining 90 percent, it was an increasing linear function. For term "low", the negation of the "high" membership function was assigned. Term Medium was the minimum of the membership degree values for term "low" and term "high". Due to the patterns/differences that occur in Scientometrics for the given linguistic variables and disciplines, each linguistic variable was modeled for each discipline separately. Two characteristic points were required to establish the membership function. The first point was the minimum for a given term (for Citation, FWCI 4y, FWCI 5y, FWCI NoWindow it was assumed to be 0, for Percentile and Team size it was assumed to be 1). To determine the values of the 90th percentile for linguistic variables, calculations were performed on a set of Scopus publications from the ICSR Lab platform (Tab. 2.).

D. Evaluation metrics

The most popular evaluation metrics used in classification algorithms were used to compare the sets obtained using the base approach and the approach using intelligent counting: Accuracy, Precision, Specificity, F1 and Matthew's correlation coefficient (MCC). The MCC metric was used due to the multi-class nature of classification, which in these cases provides a better measure of quality than Accuracy. The MlLib library for PySpark (MulticlassClassificationEvaluator class) was used to calculate the above metrics. The metrics Accuracy, Precision, Specificity, F1 were available as attributes and due to its multi-class classification nature represent their weighted average score. Due to the limitations of the MlLib library and the large dataset, the determination of MCC was based on the Macro-Averaging method for which the values TP, TN, FP, FN (also obtained by the library) and the equation for binary classification have been used. As the sets that were the subject of comparison, the results from the base approach were always used as the first set, and as the second set, respectively, each subsequent result from the application of each cardinality pattern function. Only subsets in which both the base approach and the intelligent counting approach succeeded in assigning one of the 26 classes (ignoring null values in both the first and second sets) were selected for calculation of evaluation metrics.

A supplemental measure (Set Increase) has been added to determine the percentage of observations received relative to the base approach. This measure accounted for the percentage of difference between the number of unambiguously classified observations using the intelligent counting approach and the number of unambiguously classified observations using the baseline approach to the number of observations obtained using the baseline approach (N=24,938,113). In other words, by how many percent more or less observations were successfully classified by the chosen approach than by using the base approach.

D. Classes (scientific disciplines)

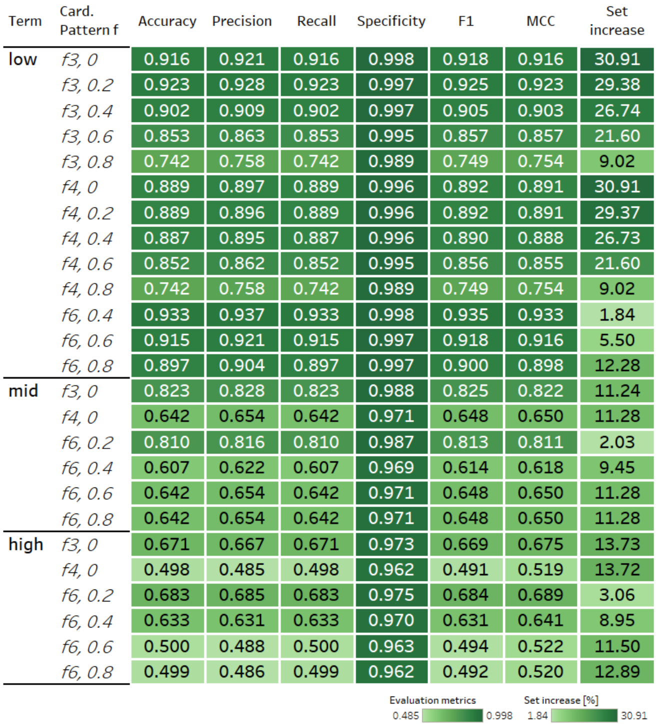
The assignment of authors to scientific disciplines was based on the ASJC (All Science Journal Classification). 26 classes were used for the study. There were 27 disciplines in the ASJC listing, but the Multidisciplinary class was excluded due to the fact that it is not a scientific discipline by itself. The following classes were used for the study: AGRI, agricultural and biological sciences; ARTS, arts and humanities; BIOC, biochemistry, genetics, and molecular biology; BUSI, business, management and accounting; CENG, chemical engineering; CHEM, chemistry; COMP, computer science; DECI, decision sciences; DENT, dentistry; EART, earth and planetary sciences; ECON, economics, econometrics and finance; ENER, energy; ENGI, engineering; ENVIR, environmental science; HEAL, health professions; IMMU, immunology and microbiology; MATE, materials science; MATH, mathematics; MEDI medicine, NEURO, neuroscience; NURS, nursing; PHARM, pharmacology, toxicology, and pharmaceutics; and PHYS, physics and astronomy, PSYC, psychology; SOCI, social sciences; VETE, veterinary.

IV. Results

In this section result has been discussed only for these Cardinality pattern functions where the number of the unambiguously assigned authors was bigger than using base approach. For Card. Pattern function $f_{5,t}$ no increase has been noted for every linguistic variable. The full results with negative Set Increase is presented on GitHub:
(link: https://github.com/lukaszszy/fedcsis-evaluation-cardinality-pattern-functions-authors-dominant-discipline-classification).

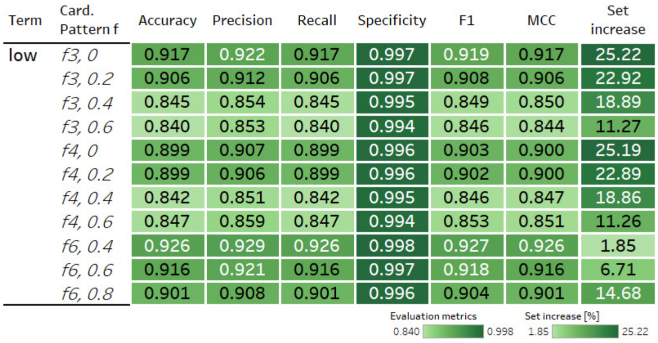
For the linguistic variable "Citation," a set increase was noted for all three terms, with the most for the “low” term. For f_3 and f_4, set increase was seen for each threshold. For f_6 only for limits 0.4 to 0.8. For term “mid” and “high” for f_3 and f_4, an increased number of observations was noted only when there was no threshold application. A significant difference can be observed between Accuracy and MCC for the “low” and “high” term. In the case of favoritism for highly cited papers, on average, more than 50 percent of researchers have classified another discipline ($f_{4,0.0}$, $f_{6,0.6}$ and $f_{6,0.8}$); giving a set increase of about 12-13 percent. The term “low” received the largest set increase of more than 30 percent, and results above 20 percent were obtained by 8 of the 13 cardinality pattern functions (Fig. 2). This is explained by the fact that there are more researchers and publications with low citations in the Scopus database than researchers and publications with a high number of citations (on the skewed distribution of publications and citations, see Albarrán et al. [2]; Carrasco and Ruiz-Castillo [6]; Ruiz-Castillo and Costas [20]).

For the FWCI metric, the number of terms for which positive set increases were achieved was incremental. For the "FWCI 4y" variable, positive results were achieved only for the low term, for "FWCI 5y" positive results were achieved for low and “mid”. For "FWCI NoWindow," positive results were achieved for all three terms. For all three FWCI variants, Accuracy and MCC results showed that classification was more similar to base classified disciplines than for the Citation variable (Dominant values above 0.8 and several Cardinality Pattern functions with result above 0.7). The largest set increase (25.22%) was recorded for FWCI 4y AND function f3,0.0. Slightly more observations (26.44%) were noted for the choice of FWCI 5y variable. For FWCI NoWindow, the number of observations over the number of observations in the baseline approach was the highest (30.30% for f4,0.0) (Fig 3., Fig 4. Fig 5.). The prerequisite for obtaining a greater variation in Sigma f-Count values for the 3 time restrictions for the FWCI indicator is the possibility of gathering a larger citation summation; as an additional 1 year for FWCI 5y compared to FWCI 4y and unlimited time for gathering citations in the case of no restriction for years (see Baas et al. [3]; de Moya-Anegón et al. [10]).



Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	f3, 0	0.916	0.921	0.916	0.998	0.918	0.916	30.91
	f3, 0.2	0.923	0.928	0.923	0.997	0.925	0.923	29.38
	f3, 0.4	0.902	0.909	0.902	0.997	0.905	0.903	26.74
	f3, 0.6	0.853	0.863	0.853	0.995	0.857	0.857	21.60
	f3, 0.8	0.742	0.758	0.742	0.989	0.749	0.754	9.02
	f4, 0	0.889	0.897	0.889	0.996	0.892	0.891	30.91
	f4, 0.2	0.889	0.896	0.889	0.996	0.892	0.891	29.37
	f4, 0.4	0.887	0.895	0.887	0.996	0.890	0.888	26.73
	f4, 0.6	0.852	0.862	0.852	0.995	0.856	0.855	21.60
	f4, 0.8	0.742	0.758	0.742	0.989	0.749	0.754	9.02
	f6, 0.4	0.933	0.937	0.933	0.998	0.935	0.933	1.84
	f6, 0.6	0.915	0.921	0.915	0.997	0.918	0.916	5.50
	f6, 0.8	0.897	0.904	0.897	0.997	0.900	0.898	12.28
mid	f3, 0	0.823	0.828	0.823	0.988	0.825	0.822	11.24
	f4, 0	0.642	0.654	0.642	0.971	0.648	0.650	11.28
	f6, 0.2	0.810	0.816	0.810	0.987	0.813	0.811	2.03
	f6, 0.4	0.607	0.622	0.607	0.969	0.614	0.618	9.45
	f6, 0.6	0.642	0.654	0.642	0.971	0.648	0.650	11.28
	f6, 0.8	0.642	0.654	0.642	0.971	0.648	0.650	11.28
high	f3, 0	0.671	0.667	0.671	0.973	0.669	0.675	13.73
	f4, 0	0.498	0.485	0.498	0.962	0.491	0.519	13.72
	f6, 0.2	0.683	0.685	0.683	0.975	0.684	0.689	3.06
	f6, 0.4	0.633	0.631	0.633	0.970	0.631	0.641	8.95
	f6, 0.6	0.500	0.488	0.500	0.963	0.494	0.522	11.50
	f6, 0.8	0.499	0.486	0.499	0.962	0.492	0.520	12.89

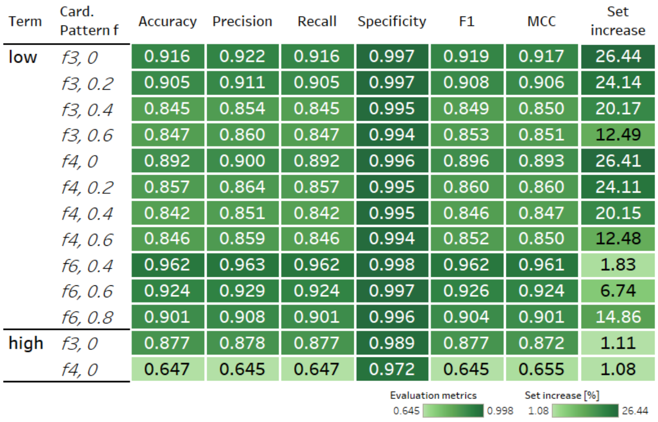
Figure 2. The results of the evaluation metrics of the linguistic variable “Citation” for the Cardinality pattern functions, where the number of uniquely identified observations exceeded the number of observations from the base approach.



Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	f3, 0	0.917	0.922	0.917	0.997	0.919	0.917	25.22
	f3, 0.2	0.906	0.912	0.906	0.997	0.908	0.906	22.92
	f3, 0.4	0.845	0.854	0.845	0.995	0.849	0.850	18.89
	f3, 0.6	0.840	0.853	0.840	0.994	0.846	0.844	11.27
	f4, 0	0.899	0.907	0.899	0.996	0.903	0.900	25.19
	f4, 0.2	0.899	0.906	0.899	0.996	0.902	0.900	22.89
	f4, 0.4	0.842	0.851	0.842	0.995	0.846	0.847	18.86
	f4, 0.6	0.847	0.859	0.847	0.994	0.853	0.851	11.26
	f6, 0.4	0.926	0.929	0.926	0.998	0.927	0.926	1.85
	f6, 0.6	0.916	0.921	0.916	0.997	0.918	0.916	6.71
	f6, 0.8	0.901	0.908	0.901	0.996	0.904	0.901	14.68

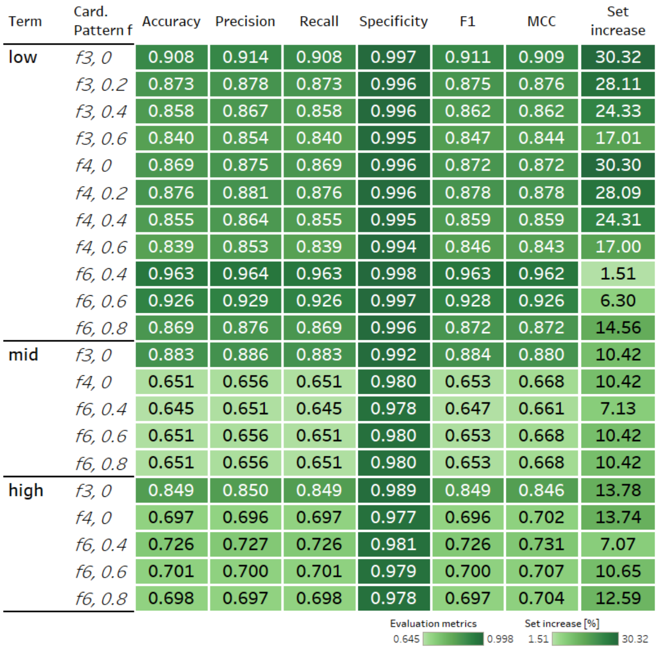
Figure 3. The results of the evaluation metrics of the linguistic variable “FWCI 4y” for the Cardinality pattern functions, where the number of

uniquely identified observations exceeded the number of observations from the base approach.



Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	f3, 0	0.916	0.922	0.916	0.997	0.919	0.917	26.44
	f3, 0.2	0.905	0.911	0.905	0.997	0.908	0.906	24.14
	f3, 0.4	0.845	0.854	0.845	0.995	0.849	0.850	20.17
	f3, 0.6	0.847	0.860	0.847	0.994	0.853	0.851	12.49
	f4, 0	0.892	0.900	0.892	0.996	0.896	0.893	26.41
	f4, 0.2	0.857	0.864	0.857	0.995	0.860	0.860	24.11
	f4, 0.4	0.842	0.851	0.842	0.995	0.846	0.847	20.15
	f4, 0.6	0.846	0.859	0.846	0.994	0.852	0.850	12.48
	f6, 0.4	0.962	0.963	0.962	0.998	0.962	0.961	1.83
	f6, 0.6	0.924	0.929	0.924	0.997	0.926	0.924	6.74
	f6, 0.8	0.901	0.908	0.901	0.996	0.904	0.901	14.86
high	f3, 0	0.877	0.878	0.877	0.989	0.877	0.872	1.11
	f4, 0	0.647	0.645	0.647	0.972	0.645	0.655	1.08

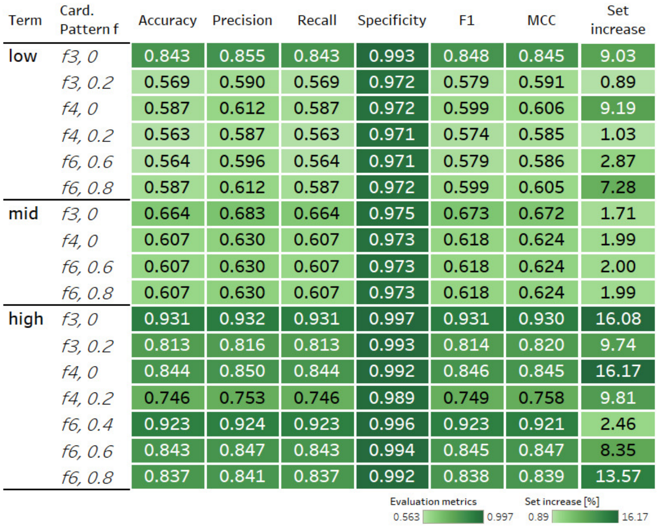
Figure 4. The results of the evaluation metrics of the linguistic variable “FWCI 5y” for the Cardinality pattern functions, where the number of uniquely identified observations exceeded the number of observations from the base approach.



Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	f3, 0	0.908	0.914	0.908	0.997	0.911	0.909	30.32
	f3, 0.2	0.873	0.878	0.873	0.996	0.875	0.876	28.11
	f3, 0.4	0.858	0.867	0.858	0.996	0.862	0.862	24.33
	f3, 0.6	0.840	0.854	0.840	0.995	0.847	0.844	17.01
	f4, 0	0.869	0.875	0.869	0.996	0.872	0.872	30.30
	f4, 0.2	0.876	0.881	0.876	0.996	0.878	0.878	28.09
	f4, 0.4	0.855	0.864	0.855	0.995	0.859	0.859	24.31
	f4, 0.6	0.839	0.853	0.839	0.994	0.846	0.843	17.00
	f6, 0.4	0.963	0.964	0.963	0.998	0.963	0.962	1.51
	f6, 0.6	0.926	0.929	0.926	0.997	0.928	0.926	6.30
	f6, 0.8	0.869	0.876	0.869	0.996	0.872	0.872	14.56
mid	f3, 0	0.883	0.886	0.883	0.992	0.884	0.880	10.42
	f4, 0	0.651	0.656	0.651	0.980	0.653	0.668	10.42
	f6, 0.4	0.645	0.651	0.645	0.978	0.647	0.661	7.13
	f6, 0.6	0.651	0.656	0.651	0.980	0.653	0.668	10.42
	f6, 0.8	0.651	0.656	0.651	0.980	0.653	0.668	10.42
high	f3, 0	0.849	0.850	0.849	0.989	0.849	0.846	13.78
	f4, 0	0.697	0.696	0.697	0.977	0.696	0.702	13.74
	f6, 0.4	0.726	0.727	0.726	0.981	0.726	0.731	7.07
	f6, 0.6	0.701	0.700	0.701	0.979	0.700	0.707	10.65
	f6, 0.8	0.698	0.697	0.698	0.978	0.697	0.704	12.59

Figure 5. The results of the evaluation metrics of the linguistic variable “FWCI NoWindow” for the Cardinality pattern functions, where the number of uniquely identified observations exceeded the number of observations from the base approach.

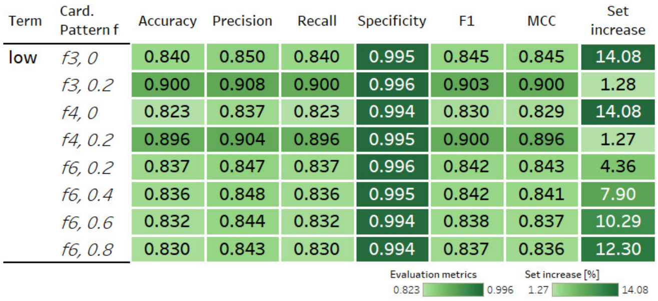
For the linguistic variable "Team size", the highest set increase was recorded (more than 16% for $f_{3,0.0}$) in the case of term “high” receiving MCC from 0.758 to 0.93. This is the opposite situation to the 4 variables mentioned above. When favoring collaborative publications, in the greatest teams, the classified discipline largely overlaps with the disciplines classified using the base approach. This fact is due that more than 90% of researchers publish only in collaborative teams, rarely having publications written solo (in STEMM disciplines, see full data on collaboration patterns across Europe by discipline in Kwiek 2021 [17]; see also Wagner and Leydesdorff [25]; Wagner [24]; Kamalski and Plume [15]). In the case of term “low” most (except $f_{3,0.0}$) cardinality pattern functions returned Accuracy and MCC scores in the range of 0.585 - 0.606, which means that even on average 40% of authors receive different scientific discipline when favoring publication in the smallest teams (Figure 6.).



Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	f3, 0	0.843	0.855	0.843	0.993	0.848	0.845	9.03
	f3, 0.2	0.569	0.590	0.569	0.972	0.579	0.591	0.89
	f4, 0	0.587	0.612	0.587	0.972	0.599	0.606	9.19
	f4, 0.2	0.563	0.587	0.563	0.971	0.574	0.585	1.03
	f6, 0.6	0.564	0.596	0.564	0.971	0.579	0.586	2.87
	f6, 0.8	0.587	0.612	0.587	0.972	0.599	0.605	7.28
mid	f3, 0	0.664	0.683	0.664	0.975	0.673	0.672	1.71
	f4, 0	0.607	0.630	0.607	0.973	0.618	0.624	1.99
	f6, 0.6	0.607	0.630	0.607	0.973	0.618	0.624	2.00
	f6, 0.8	0.607	0.630	0.607	0.973	0.618	0.624	1.99
high	f3, 0	0.931	0.932	0.931	0.997	0.931	0.930	16.08
	f3, 0.2	0.813	0.816	0.813	0.993	0.814	0.820	9.74
	f4, 0	0.844	0.850	0.844	0.992	0.846	0.845	16.17
	f4, 0.2	0.746	0.753	0.746	0.989	0.749	0.758	9.81
	f6, 0.4	0.923	0.924	0.923	0.996	0.923	0.921	2.46
	f6, 0.6	0.843	0.847	0.843	0.994	0.845	0.847	8.35
	f6, 0.8	0.837	0.841	0.837	0.992	0.838	0.839	13.57

Figure 6. The results of the evaluation metrics of the linguistic variable “Team size” for the Cardinality pattern functions, where the number of uniquely identified observations exceeded the number of observations from the base approach.

The least number of cardinality pattern functions that returned a positive set increase were obtained for the variable "Percentile". The results show that Accuracy and MCC for each function is more than 0.83. In addition, it can be noted that for function f_6, the greater the favoritism of publications in the lower percentile of the journal, the greater the set increase becomes (from 4.36 to 12.3 which is about 3 times additional observations). (Fig. 7). This situation is explained by the fact that in science there are many scientists who tend to submit their papers to journals with a lower prestige (see Blackmore and Kandiko [4]; Franzoni et al.[12]; Starbuck [22]).



Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	f3, 0	0.840	0.850	0.840	0.995	0.845	0.845	14.08
	f3, 0.2	0.900	0.908	0.900	0.996	0.903	0.900	1.28
	f4, 0	0.823	0.837	0.823	0.994	0.830	0.829	14.08
	f4, 0.2	0.896	0.904	0.896	0.995	0.900	0.896	1.27
	f6, 0.2	0.837	0.847	0.837	0.996	0.842	0.843	4.36
	f6, 0.4	0.836	0.848	0.836	0.995	0.842	0.841	7.90
	f6, 0.6	0.832	0.844	0.832	0.994	0.838	0.837	10.29
	f6, 0.8	0.830	0.843	0.830	0.994	0.837	0.836	12.30

Figure 7. The results of the evaluation metrics of the linguistic variable “Percentile” for the Cardinality pattern functions, where the number of uniquely identified observations exceeded the number of observations from the base approach.

V. SUMMARY

For the linguistic variables presented above, it can be seen that the application of intelligent counting contributes significantly to increasing the number of unambiguously classified disciplines to authors. For selected Cardinality patter functions it was possible to expect a return that exceeds by 25% the number of results using the base approach (about 30% for Citation and FWCI NoWindow, 26% for FWCI 5y, 25 for FWCI 4y; in all cases for term "low"). One has to wonder, however, whether it is reasonable to use the term "low", if we would like to base our research on the least cited publications and on publications with the lowest FWCI metric. What would be reasonable in this case is to operate on the term "high" and the satisfaction of increasing the number of observations by 13-16% (around 16% for Team Size and 13% for Citation and FWCI NoWindow).

The results presented above provide a basis for further analysis of the presented problem. It is necessary to focus on further unexplored scientific metrics and cardinality pattern functions to examine their influence on the determination of the dominant discipline. Due to the large number (31%) of authors who received assignment to more than one dominant discipline, it would be interesting to consider a multi-label classification variant as an alternative to multi-class classification. Besides discipline, there are other locations where the conceptual apparatus of fuzzy logic can be applied. A dimension that also needs to be explored is the author's dominant country or their full affiliation.

The other approach that may also prove interesting assumes is the use of fuzzy controllers [16]. Appropriate definition of a set of rules for a fuzzy controller can also set a new direction, and thus hypothetically improve the algorithm presented in the classical approach. Besides the scientific database Scopus, there are other databases: Web of Science, OpenAlex (or subsets of these databases) for which the above algorithm can be applied. Therefore, the next step in the work on the algorithm may be to integrate or compare results from many different data sources and applying a voting mechanism to them [14].

APPENDIX

TABLE II.
VALUES OF TOP 10TH PERCENTILE FOR MEMBERSHIP FUNCTIONS BY DISCIPLINES

Discipline	Citation	FWCI 4y	FWCI 5y	FWCI NoWindow	Team size	Percentile	Number of publications
AGRI	53	2.36	2.34	2.33	7	53	4,827,009
ARTS	26	2.59	2.53	2.49	3	26	2,144,574
BIOC	74	2.55	2.53	2.46	9	74	8,716,173
BUSI	54	2.98	2.96	2.91	4	54	1,219,375
CENG	66	2.94	2.92	2.88	7	66	2,855,087
CHEM	60	2.67	2.64	2.57	7	60	6,582,601
COMP	45	2.84	2.81	2.68	5	45	2,835,794
DECI	56	2.95	2.95	2.95	4	56	490,649
DENT	45	2.42	2.40	2.38	7	45	405,592
EART	55	2.49	2.47	2.35	7	55	2,800,145
ECON	46	2.75	2.70	2.60	4	46	929,703
ENER	53	3.09	3.08	3.04	7	53	1,611,907
ENGI	40	2.83	2.79	2.62	6	40	9,031,637
ENVI	55	2.79	2.76	2.69	7	55	3,432,530
HEAL	43	2.62	2.59	2.52	7	43	724,899
IMMU	73	2.57	2.54	2.48	10	73	2,037,654
MATE	50	2.80	2.78	2.66	7	50	5,973,127
MATH	35	2.48	2.44	2.32	4	35	3,112,501
MEDI	44	2.19	2.19	2.21	8	44	20,632,131
NEUR	80	2.57	2.55	2.47	9	80	1,690,716
NURS	40	2.59	2.56	2.47	7	40	1,059,202
PHAR	47	2.33	2.31	2.24	8	47	2,337,260
PHYS	51	2.70	2.68	2.50	7	51	7,442,223
PSYC	64	2.64	2.63	2.58	6	64	1,511,213
SOCI	34	2.60	2.56	2.46	4	34	4,492,180
VETE	33	2.46	2.40	2.31	8	33	586,876

ACKNOWLEDGMENT

I gratefully acknowledge the support of my supervisor from Centre for Public Policy Studies Prof. Marek Kwiek. I also gratefully acknowledge the assistance of the International Center for the Studies of Research (ICSR Lab) for research purposes and Kristy James, Senior Data Scientist in the ICSR Lab, for her continuous support.

AUTHOR CONTRIBUTIONS

Lukasz Szymula - Conceptualization, Data curation, Formal Analysis, Methodology, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing.

Krzysztof Dyczkowski - Conceptualization, Formal Analysis, Methodology, Validation, Writing - original draft, Writing - review & editing.

COMPETING INTERESTS

No potential conflict of interest was reported by the author(s).

OPEN ACCESS PRACTICES

Access to the Scopus database was granted through a collaboration between AMU Center for Public Policy Studies and the International Center for the Study of Research (ICSR Lab), Elsevier, established in November 2020 so the dataset is not publicly available. Full results of the study are available on GitHub:
(link: https://github.com/lukaszszy/fedcsis-evaluation-cardinality-pattern-functions-authors-dominant-discipline-classification).

REFERENCES

[1] Abramo, G., Aksnes D. W., D'Angelo C. A., 2020. Comparison of research productivity of Italian and Norwegian professors and universities. Journal of Informetrics, 14(2), 101023.

[2] Albarrán, P., Crespo, J. A., Ortuño, I., Ruiz-Castillo, J., 2011. The skewness of science in 219 sub-fields and a number of aggregates. Scientometrics. Vol. 88(2). 385–397.

[3] Baas, J., Schotten, M., Plume, A., Côté, G., Karimi, R., 2020. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies. Vol. 1(1). 377–386. https://doi.org/10.1162/qss_a_00019.

[4] Blackmore, P., & Kandiko, C. B., 2011. Motivation in academic life: A prestige economy. Research in Post-Compulsory Education, 16(4), 399–411.

[5] Boekhout, H., van der Weijden, I., Waltman L., 2022. Gender differences in scientific careers: A large-scale bibliometric analysis . Available from: https://arxiv.org/abs/2106.12624

[6] Carrasco, R., Ruiz-Castillo, J., 2014. The evolution of the scientific productivity of highly productive economists. Economic Inquiry. Vol. 52(1). 1–16.

[7] Cassidy R.S., Larivièrev., 2018. Measuring Research, What Everyone Needs to Know. Oxford University Press.

[8] Chan, H. and Torgler, B., 2020. Gender differences in performance of top cited scientists by field and country. Scientometrics, 125(3), pp.2421-2447, https://doi.org/10.1007/s11192-020-03733-w.

[9] Daradkeh M, Abualigah L, Atalla S, Mansoor W., 2022. Scientometric Analysis and Classification of Research Using Convolutional Neural Networks: A Case Study in Data Science and Analytics. Electronics; 11(13):2066. https://doi.org/10.3390/electronics11132066

[10] de Moya-Anegón, F., Z. Chinchilla-Rodríguez, B. Vargas-Quesada, E. Corera-Álvarez, F. Munoz-Fernández, A. Gonzalez-Molina, and V. Herrero-Solana., 2007. Coverage Analysis of Scopus: A Journal Metric Approach. Scientometrics 73 (1): 53–78.

[11] Dyczkowski K., 2018. Intelligent Medical Decision Support System Based on Imperfect Information. The Case of Ovarian Tumor Diagnosis. Studies in Computational Intelligence, https://doi.org/10.1007/978-3-319-67005-8.

[12] Franzoni, C., Scellato, G., & Stephan, P., 2011. Changing incentives to publish. Science, 333(6043), 702–703.

[13] Jöns, H., Hoyler, M., 2013. Global geographies of higher education: The perspective of world university rankings. Geoforum, 46, pp.45-59, https://doi.org/10.1016/j.geoforum.2012.12.014.

[14] Kacprzyk, J., 1985. Group decision-making with a fuzzy majority via linguistic quantifiers. Part I: a consensory-like pooling. Cybernetics and Systems, 16(2-3), pp.119-129, https://doi.org/10.1080/01969728508927761.

[15] Kamalski, J., and Plume. A., 2013. Comparative Benchmarking of European and US Research Collaboration and Researchers Mobility: A Report Prepared in Collaboration Between Science Europe and Elsevier's SciVal Analytics. Science Europe, Elsevier.

[16] Kickert, W., Mamdani, E., 1978. Analysis of a fuzzy logic controller. Fuzzy Sets and Systems, 1(1), pp.29-44, https://doi.org/10.1016/0165-0114(78)90030-1.

[17] Kwiek, M., 2021. What large-scale publication and citation data tell us about international research collaboration in Europe: Changing national patterns in global contexts. Studies in Higher Education. Vol. 46(12). 2629–2649.

[18] Kwiek, M., Roszka, W., 2022. Academic vs. biological age in research on academic careers: a large-scale study with implications for scientifically developing systems. Scientometrics, https://doi.org/10.1007/s11192-022-04363-0.

[19] Meen C. K., Seojin N., Fei W., Yongjun Z., 2020. Mapping scientific landscapes in UMLS research: A scientometric review. Journal of the American Medical Informatics Association, 27(10), 1612–1624, https://doi.org/10.1093/jamia/ocaa107

[20] Ruiz-Castillo, J., Costas, R., 2014. The skewness of scientific productivity. Journal of Informetrics. Vol. 8(4). 917–934.

[21] Sood, S.K., Kumar, N. & Saini, M., 2021. Scientometric analysis of literature on distributed vehicular networks : VOSViewer visualization techniques. Artif Intell Rev 54, 6309–6341, https://doi.org/10.1007/s10462-021-09980-4

[22] Starbuck, W. H., 2013. Why and where do academic publish? M@n@gement, 5, 707–718.

[23] Visser, M., van Eck, N. and Waltman, L., 2021. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies, 2(1), pp.20-41, https://doi.org/10.1162/qss_a_00112.

[24] Wagner, C. S., 2018. The Collaborative Era in Science. Governing the Network. Cham: Palgrave Macmillan.

[25] Wagner, C. S., and L. Leydesdorff., 2005. Network Structure, Self-Organization, and the Growth of International Collaboration in Science. Research Policy 34 (10): 1608–18, https://doi.org/10.1016/j.respol.2005.08.002.

[26] Wygralak M., 2015. Intelligent Counting Under Information Imprecision. Applications to Intelligent Systems and Decision Support. , Studies in Fuzziness and Soft Computing, https://doi.org/10.1007/978-3-642-34685-9.

[27] Zadeh L.A., 1975. The concept of a linguistic variable and its application to approximate reasoning—I. Information Sciences, Volume 8, Issue 3, https://doi.org/10.1007/978-1-4684-2106-4_1.