Dengesiz Veri Kümeleriyle Sınıflandırmada Gelişen Trendler: İlerlemenin Bibliyometrik Analizi

Araştırma Makalesi/Research Article

Abdullah MARAŞ[1], Çiğdem SELÇUKCAN EROL[2]

[1]İstanbul Üniversitesi Fen Bilimleri Enstitüsü Enformatik Ana Bilim Dalı, İstanbul, Türkiye
[2] İstanbul Üniversitesi Enformatik Bölümü ve Fen Fakültesi Biyoloji Bölümü Botanik Ana Bilim Dalı, İstanbul, Türkiye
abdullahmaras@ogr.iu.edu.tr, cigdems@istanbul.edu.tr

(Geliş/Received:04.11.2021; Kabul/Accepted:06.06.2022)
DOI: 10.17671/gazibtd.1019015

Özet— Dengesiz veri kümeleri, makine öğrenimi alanında hedef değişkenin oldukça çarpık dağılımı olarak tanımlanmaktadır. Dengesiz veri kümeleri, makine öğrenimi modelleri üzerindeki olumsuz etkilerinden dolayı son on yılda araştırmacıların dikkatini büyük ölçüde çekmiştir. Araştırmacılar dengesiz veri kümeleri sorunlarına çeşitli çözümler geliştirip literatürde paylaşmaktadır.

Artan makale sayısı literatürü takip etmeyi zorlaştırmaktadır. Derleme makaleleri bu sorunun çözümüne katkıda bulunur. Bu çalışmada, dengesiz veri kümeleriyle sınıflandırmadaki çözüm önerilerini bulmak için bibliyometrik bir analiz yapılması amaçlanmaktadır. Bibliyometrik analiz, veri tabanlarından istatistik çıkarmaya dayalı nicel bir tekniktir. Bu çalışma, dengesiz veri kümeleri problemini ele alan ilk bibliyometrik analizi olma niteliğindedir.

Bu çalışmada, Scopus veri tabanından, dengesiz veri kümeleriyle ilgili veri, R Bibliometrix package version 3.1.4 ile elde edilerek son çalışmalar ve yeni yaklaşımlar özetlendi. Seçilen anahtar kelimeler ile 1957-2021 yılları arasında 16255 yayına ilişkin veriler toplandı. Bu koleksiyon temel olarak 8871 makale, 6987 konferans bildirisi ve 175 derlemeden oluşmaktadır ve belge başına atıf sayısı yılda ortalama 1,66'dır. En çok atıf yapılan ülkeler arasında 106139 toplam atıf ile Amerika Birleşik Devletleri'ni, 13839 atıf ile Çin ve 9524 atıf ile Almanya takip etmektedir.

Anahtar Kelimeler— dengesiz öğrenim, sınıflandırma, örnekleme yöntemleri, maliyet duyarlı çözüm, değerlendirme metrikleri, bibliyometrik

Emerging Trends in Classification with Imbalanced Datasets: A Bibliometric Analysis of Progression

Abstract— Imbalanced or unbalanced datasets are defined as the highly skewed distribution of target variable in the field of machine learning. Imbalanced datasets have greatly caught the attention of researchers due to their negative effect on machine learning models in the last decade. Researchers develop various solutions to the problems of imbalanced datasets and contribute to the literature.

The increasing number of articles makes it difficult to follow the literature. Review articles contribute to the solution of this problem. The goal of this study is to conduct a bibliometric analysis to find solutions for classification with imbalanced datasets. Bibliometric analysis is a quantitative technique based on extracting statistics from databases. This work is the first bibliometric analysis to address the problem of imbalanced datasets.

In this study, data on imbalanced datasets were obtained from the Scopus database with the R Bibliometrix package version 3.1.4, and recent studies and new approaches were summarized. Data on 16255 publications between 1957-2021 were collected by using selected keywords. This collection mainly comprises 8871 articles, 6987 conference papers, and 175 reviews with 1, 66 average citations per year per document. Among the most cited countries, the United States has 106139 total citations followed by China with 13839 citations and Germany has 9524 citations.

Keywords— Imbalanced learning, classification, sampling methods, cost-sensitive learning, evaluation metrics, bibliometric

1. INTRODUCTION

Recent developments in technology opened the doors of the data age to companies so that companies have started to change the way of doing business. The rise of usage in products related to Internet of things (IoT), social media, e-commerce, online banking and many other things bring, not only the expansion of stored data, but also big advances in data science with it. The approaches to gain insight from data, starting with basics of statistics, have led various areas of study from machine learning to deep learning.

One of the most studied topics in the field is classification problems which contain a wide range of applications. Classification problems, a supervised learning approach, aims to predict one of the predefined target variable [1]. Fraud detection [2], credit scoring [3], medical diagnosis [4-6], churn prediction in various fields [7], and sentiment analysis [8] are some main examples. Numerous type of algorithms, such as random forest, decision trees, SVMs, neural nets, logistic regression, and naive bayes, have been applied to solve classification problems [9]. The models developed using these algorithms leads to successful results if the dataset has a balanced distribution of the target variable. However it is also observed that in the presence of skewed dataset, so called imbalanced dataset, the performance of the models are comparatively low [10].

Imbalanced datasets are defined as; one of the target variables mostly heavily outnumbers the other target variable [11]. A high accuracy score can be obtained in the case of imbalanced datasets since the model prone to predict the majority target variable instances [12]. However, a high accuracy score is not always an indication of good model. Although good prediction of majority target variable, the minority target variable has always not only significant importance but also difficult part of the problem to predict. For instance, it is important to predict correct fraudulent transactions for a fraud detection problem albeit fewer fraud examples [2]. Since the negative effects of imbalanced datasets on model are an undeniable fact, there is a great interest in topic that can be easily observed in google search trends in Fig. 1 [13]. The growing interest in the field obvious especially in last decade. The importance of the topic yields to two workshops which are AAAI 2000 [14] and ICML 2003 [15] conferences. From solutions to the problem to evaluation of imbalanced datasets various topics are discussed during these workshops. After these workshops the field gained momentum. Although there is a slowdown in interest up to 2014, we can observe the rise of the field after this period.

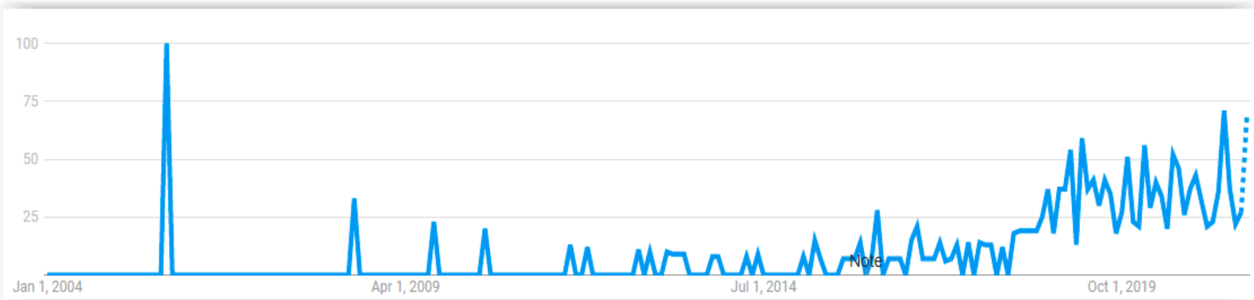

Figure 1. Imbalanced dataset search on Google

Imbalanced datasets always have different degrees of target variables skewness in daily applications. Imbalanced degree of a classification problem is called imbalanced ratio (IR). The imbalanced ratio can be defined as ratio of the majority target variable (prevalent class) to the minority target variable (small class) [11, 16-18]. In real world applications the imbalanced ratio can be as big as 100:1, 1000:1 or even bigger [11]. Imbalanced ratio has a great impact on performance of machine learning algorithms so that there are different approaches to solve this problem.

The cause of imbalanced datasets can be intrinsic or extrinsic [17]. Intrinsic imbalanced datasets problems occur due to nature of the problem. Most medical classification problems are great examples of intrinsic datasets. In this kind of datasets, there is always one target variable that has fewer examples than other variables [19]. Contrarily extrinsic imbalanced datasets can be result of insufficient data collection time or data storage. A balanced dataset problem can turn into an imbalanced dataset problem due to aforementioned reasons. Classification with streaming data is a pervasive extrinsic datasets problem [20].

Besides the effects of imbalanced datasets on model performance, there are researches on other factors that can cause problem for model performance [11, 21-23]. Small disjuncts [22, 23], class overlapping [24], and noisy examples in datasets [25, 26] are among the other factors that hinder model performance. In this research, bibliometric analysis is used to extract general understanding of imbalanced dataset problem by identifying featured studies, journals and emerging approaches.

2. METHODOLOGY

Rapid increase of machine learning related research publications can be overwhelming to follow the field. Bibliometric analysis contains quantitative and statistical analysis to overview of a research field and discover rising trends [27].

It is important to comprehend the bibliometric analysis structure before delving into result details. The methodological framework used in this study is presented in Fig. 2. The adopted components are represented with shaded boxes. Our study is an example of explanatory research that is described as a technique for gathering data in order to explain a situation. Our research logic is a bottom-up strategy which is also known as inductive. We investigate both context and metadata of the publications are also addressed as primary and secondary data in the figure. The research consists of data collected from Scopus [28] over the period of 1957-2021 for articles that include keywords *imbalanced dataset, imbalanced big dataset, oversampling, smote, unbalanced big dataset, unbalanced dataset* and *undersampling* . After removal of missing values and repetitions, metadata belongs to 16255 publications remained. The preprocessed data is analyzed by using R Bibliometrix package version 3.1.4 by Aria et al. [30] which is a tool for conducting bibliometric analysis. Bibliometrix package not only facilitates co-citation, coupling, collaboration, and co-word analysis but also visualizations [30].

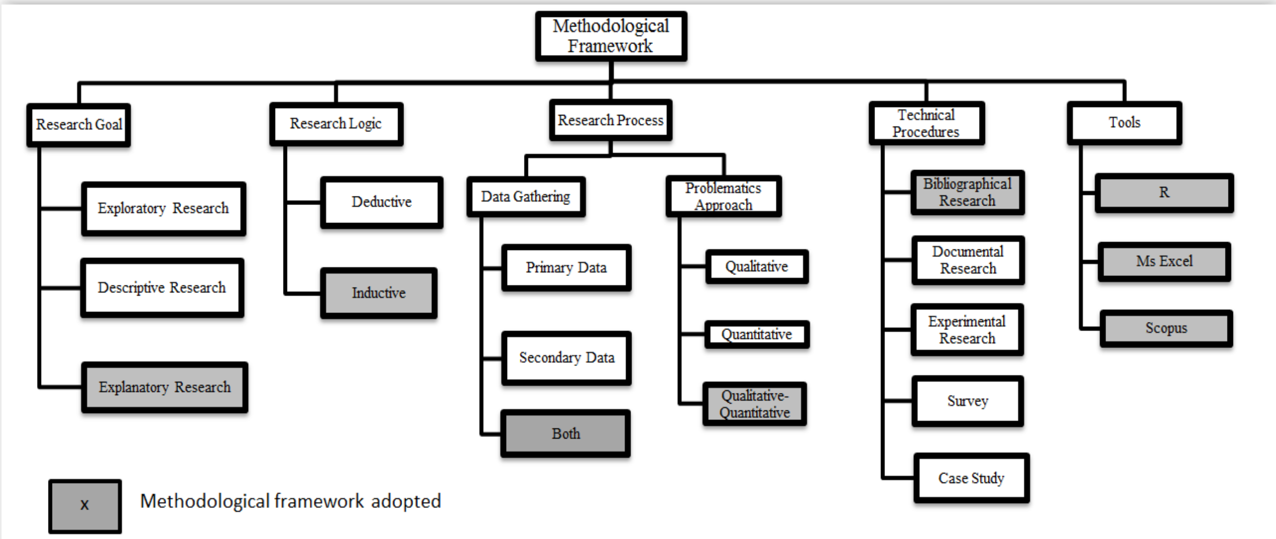

Figure 2. Methodological framework [29]

3. BIBLIOMETRIC ANALYSIS RESULTS

We give a quantitative study of 16255 papers from annual scientific production, as well as historiography and overview of the field's trend.

3.1. Annual Scientific Production

The preprocessed data were analyzed to explore the growth of the subject. Fig. 3 shows the annual number of publications between the years 1957 and 2021. Interest in imbalanced datasets seems to be very rare in the early stages of artificial intelligence research. Especially after 2003, there is global attention on the topic. The dramatic increase in the publication is very obvious in Fig. 3.

3.2. Average Total Citations

Although there is a clear interest in imbalanced datasets as demonstrated in the Fig. 3, Fig. 4 shows that average citations are decreasing and slightly fluctuating. The most average citations are observed in 2002. Following a thorough examination of publications from this year, it is considered that the peak is the result of one of the well-known unbalanced dataset problems' solutions, Synthetic Minority Over-sampling Technique (SMOTE), developed by Chawla et al. in 2002 [31]. SMOTE and its variations are still the most often used method for dealing with imbalanced datasets.

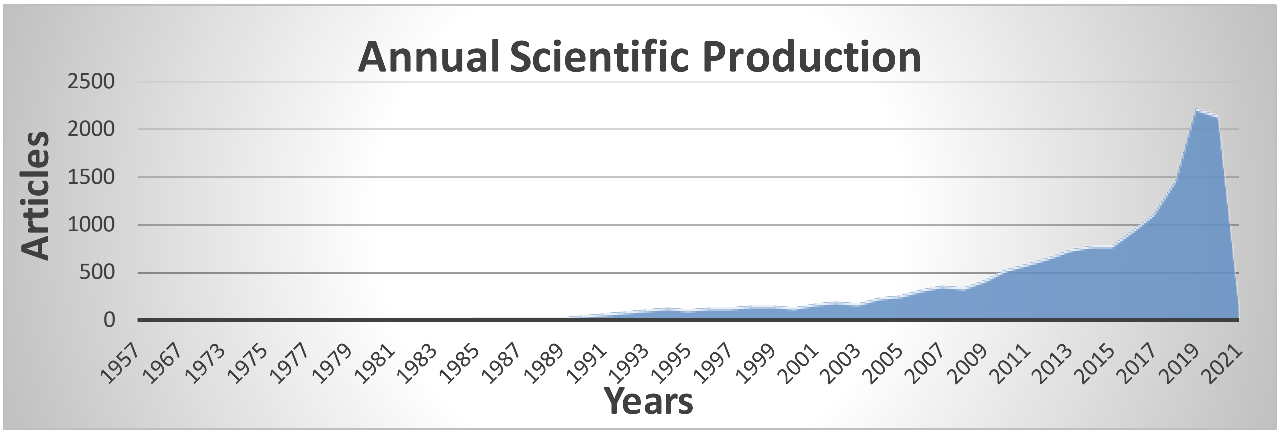

Figure 3. The Scopus publications on imbalanced datasets from 1957 to 2021

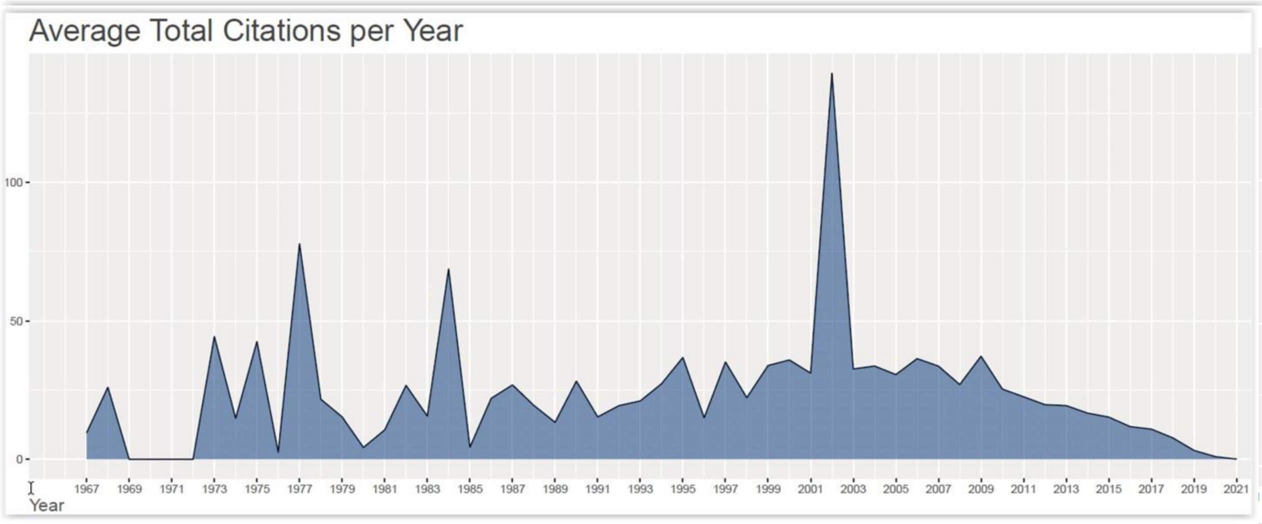

Figure 4. Average total citations per year

3.3. Country Contributions on Imbalanced Datasets

Fig. 5 illustrates the most productive countries. It is important for researchers to know the countries that contribute their field. According to Fig. 5, the majority of papers are published by authors in the United States, followed by China and Germany respectively. It can be also deduced from Fig. 5 that there are more single country publications than multiple country publications.

MCP indicates, for each country, the number of documents in which there is at least one co-author from a different country. The United States has also the most country collaborations in the field. However, when we look at the percentages (MCP/ (MCP+SCP)) of countries with high levels of international collaboration, the United Kingdom comes out on top with 32%, followed by Germany (26%), and Spain (23%) respectively.

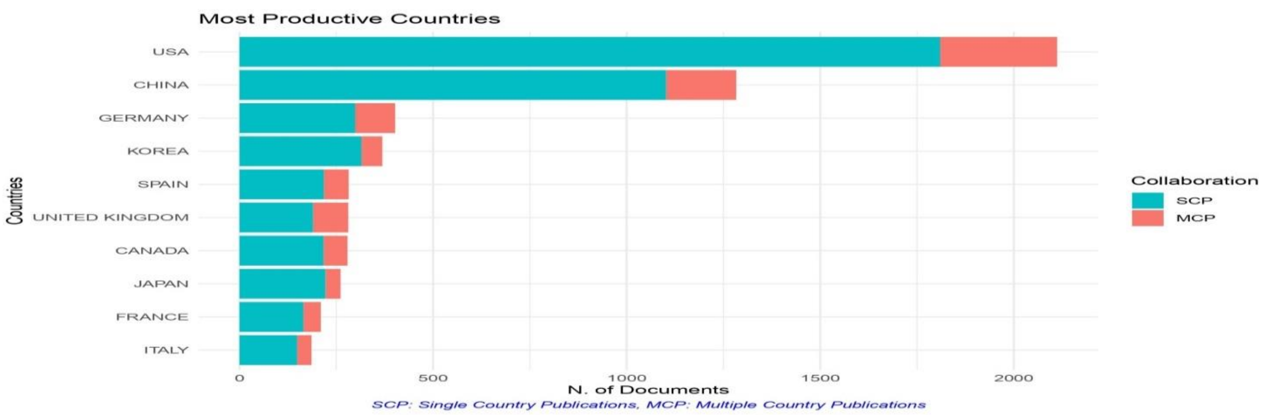

Figure 5. Countries contribution on problem

Countries collaboration and prevalence of research by countries is shown in Fig. 6. As can be seen in Fig. 6, there is substantial collaboration between the United States and China, as well as the United States and Europe, particularly Germany. Hong Kong, the United Kingdom, and Canada all seem to be strong partners for China.

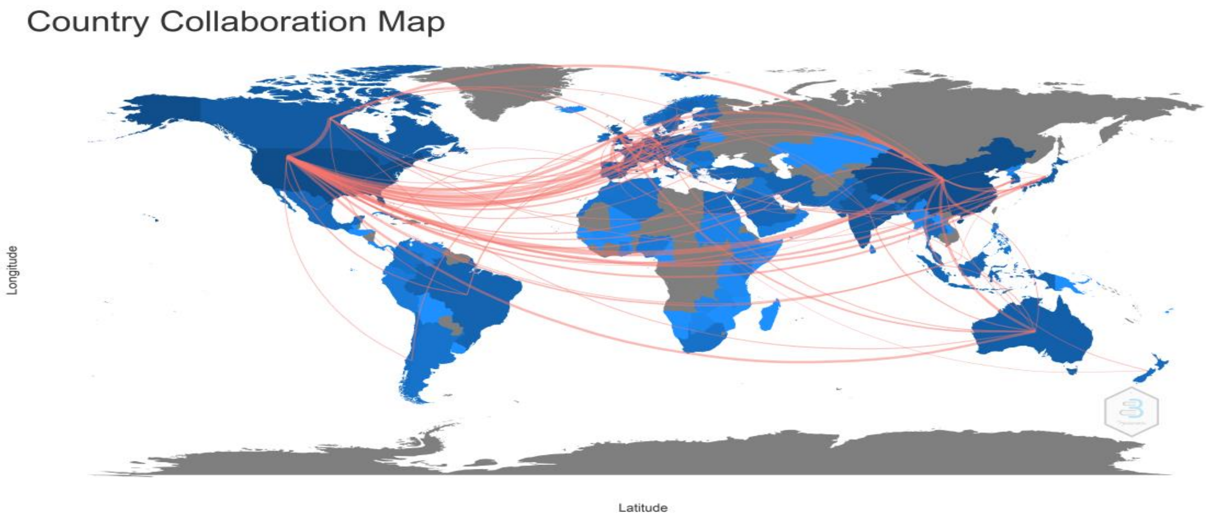

Figure 6. Collaboration of countries on world map

3.4. Highly Cited Publications

Table 1 which is obtained through the web-interface of bibliometrix package, shows the most cited 20 publications in the domain of imbalanced datasets problem. Total citations show how many times each article has been cited, whereas local citations show how many times an author (or a document) in this collection has been cited by other authors in this collection. As mentioned in section 3.2, Chawla et al. [31] study's sparked a lot of attention and it is now one of the most referenced papers in the domain. The study proposed an over-sampling approach by creating synthetic examples of minority data [31]. Borderline-SMOTE paper is the second highest cited paper that proposed a new technique on top of SMOTE [32]. Han et al. developed two new oversampling approaches, borderline-SMOTE1 and borderline-SMOTE2, that only oversample the minority examples around the borderline [32]. In comparison to SMOTE, the borderline method produces better outcomes in terms of TP rate and F-value [32]. The third highly cited paper belongs to Chawla et al. is called SMOTEBoost where they proposed a hybrid solution that combines SMOTE algorithm and the boosting procedure [33]. They achieved better success in predicting the minority class and overall F-value on severely and moderately unbalanced datasets. Galar et al. published a review on class imbalance problem [34]. According to the study, ensemble of random undersampling techniques with bagging or boosting yields decent results. It is also indicated that ensemble-based algorithms outperforms preprocessing methods. Undersampling is one of the main techniques used to handle with imbalanced datasets. Liu et al. point out the weakness of undersampling technique and proposed two new approaches to cope with it [35]. The proposed methods EasyEnsemble and BalanceCascade take also majority class examples into consideration which are overlooked in the undersampling technique. Based on the study these methods not only give the better area under the ROC curve, F-measure, and G-mean values but also the training time remains almost the same. Imbalanced datasets don't decrease degree of learning by themselves [36]. López et al. highlight subproblems of imbalanced datasets as small disjuncts, lack of density, overlapping or class separability, noisy data, borderline examples, and datasets shift [36]. They also conducted a comparative research to contrast imbalanced datasets solutions [36]. Support Vector Machines (SVM) is one of the frequently used algorithms in classification problems and also suffers from datasets imbalances. Akbani et al. [37] discuss performance degradation of SVM in the case of imbalance dataset and come up with a hybrid method to address the drawback of the undersampling technique. They introduced the SMOTE with Different Costs (SDC) approach which combines SMOTE by Chawla et al. [31] with different error costs algorithm by Veropoulos et al [38]. Another hybrid method, RUSBoost, is researched by Seiffert et al [39]. RUSBoost is a variant of SMOTEBoost that incorporates a combination of random undersampling and AdaBoost. Although both SMOTEBoost and RUSBoost techniques are based on AdaBoost algorithm, it is stated that differentiation of sampling methods makes RUSBoost faster than SMOTEBoost [39]. Zhou et al. [40] answered the questions whether cost-sensitive learning methods are also effective on solving imbalanced datasets problem and vice versa. Their research on 21 UCI datasets shows that cost-sensitive neural networks are difficult in the presence of a higher degree of class imbalance [40]. Estabrooks and Japkowicz [41] investigate effectiveness of undersampling and oversampling approach and presented a hybrid scheme to reach better results. A detailed study on combination of SVM and sampling methods is conducted by Tang et al. [42]. They proposed four approaches and compared these approaches with state-of-the-art solutions. According to their study, their Granular Support Vector Machines – Repetitive Undersampling (GSVM-RU) gives better results among four proposed approaches which aim to directly utilize SVM itself for undersampling [42]. SMOTE technique increases minority class by synthesizing new instances from closest neighbors, regardless of majority class examples. Safe-Level-SMOTE defines safe level as the number of positive instances in k nearest neighbors and aims to generate all synthetic instances in safe level [43]. According to Bunkhumpornpat et al. [43] research, Safe-Level-SMOTE achieves better F-value in comparison to SMOTE and Borderline-SMOTE. Wang and Yao [44] investigate the impact of diversity in imbalanced datasets and stated that larger diversity causes better recall for minority but worse recall for majority classes. Medical datasets are always prone to be imbalanced due to the nature of the problem. Research on the effects of imbalanced datasets in medical diagnosis shows that model performance decrease in the case of imbalances [45]. Mazurowski et al. [45] implements backpropagation (BP) and particle swarm optimization (PSO) neural networks and report that BP gives always better results than PSO with imbalanced datasets. EUSBoost based on RUSBoost is proposed by Galar et al. [46], to put more emphasis on diversity. EUSBoost results a dataset which contains all minority class examples and selected majority class examples and ensuring the diversity by training each classifier with different subsets of the majority class [46]. The publication of SMOTE technique forged great impact on imbalanced datasets literature and yields different variations of it. Fernandez et al. [47] not only analyze the studies based on SMOTE and its application but also identify challenges for SMOTE on Big Data. The growing volume of data causes performance issues when using existing methods for imbalanced datasets. Triguero et al. [48] propose a MapReduce scheme for undersampling approach to overcome performance issues. Sáez et al. [49] also underline importance of noisy and borderline examples in imbalanced datasets and recommend an extension of SMOTE to alleviate disadvantages of it. They apply Iterative-Partitioning Filter (IPF) to filter noisy examples and clean up class boundaries. Biomedical and bioinformatics are other fields that struggle with high dimensionality and imbalanced datasets. Blagus and Lusa [50] proposed Geometric Mean Nearest Shrunken Centroid (GM-NSC) which targets to maximize g-means to estimate optimal shrinkage. Imbalanced datasets problem in credit scoring is investigated by Brown and Mues [51]. According to their study, the random forests and the (Placeholder1) gradient boosting algorithms yield better results while techniques such as Quadratic discriminant analysis (QDA) and C4.5 gives worse results.

Table 1. The summary of the most cited 20 articles

Document	DOI	Local Citations	Total Citations
CHAWLA NV, 2002, J ARTIF INTELL RES	10.1613/jair.953	2915	7540
HAN H, 2005, LECT NOTES COMPUT SCI	10.1007/11538059_91	756	1131
CHAWLA NV, 2003, LECT NOTES ARTIF INTELL	10.1007/978-3-540-39804-2_12	504	834
GALAR M, 2012, IEEE TRANS SYST MAN CYBERN PT C APPL REV	10.1109/TSMCC.2011.2161285	435	1145
LIU XY, 2009, IEEE TRANS SYST MAN CYBERN PART B CYBERN	10.1109/TSMCB.2008.2007853	387	915
LPEZ V, 2013, INF SCI	10.1016/j.ins.2013.07.007	336	704
AKBANI R, 2004, LECT NOTES ARTIF INTELL	10.1007/978-3-540-30115-8_7	329	713
SEIFFERT C, 2010, IEEE TRANS SYST MAN CYBERN PT A SYST HUMANS	10.1109/TSMCA.2009.2029559	320	734
ZHOU ZH, 2006, IEEE TRANS KNOWL DATA ENG	10.1109/TKDE.2006.17	293	732
ESTABROOKS A, 2004, COMPUT INTELL	10.1111/j.0824-7935.2004.t01-1-00228.x	249	556
TANG Y, 2009, IEEE TRANS SYST MAN CYBERN PART B CYBERN	10.1109/TSMCB.2008.2002909	215	498
BUNKHUMPORNPAT C, 2009, LECT NOTES COMPUT SCI	10.1007/978-3-642-01307-2_43	213	324
WANG S, 2009, IEEE SYMP COMPUT INTELL DATA MIN , CIDM - PROC	10.1109/CIDM.2009.4938667	142	254
MAZUROWSKI MA, 2008, NEURAL NETW	10.1016/j.neunet.2007.12.031	141	419
GALAR M, 2013, PATTERN RECOGN	10.1016/j.patcog.2013.05.006	119	198
FERNNDEZ A, 2018, J ARTIF INTELL RES	10.1613/jair.1.11192	106	163
GARCA S, 2009, EVOL COMPUT	10.1162/evco.2009.17.3.275	104	218
SEZ JA, 2015, INF SCI	10.1016/j.ins.2014.08.051	102	183
BLAGUS R, 2013, BMC BIOINFORM	10.1186/1471-2105-14-106	90	191
BROWN I, 2012, EXPERT SYS APPL	10.1016/j.eswa.2011.09.033	87	272

3.5. The Most Common Keywords

The analysis of publications keywords is great and important step to get first insight about the field. Table 2 provides the some of the most used keywords by authors. Based on the keywords it is obvious that authors focus on balancing datasets by using one of the sampling methods. Also, oversampling methods attract the author's attention the most among sampling methods. Posterior to the sampling methods authors focused on algorithms which are random forest, support vector machine and ensemble models. It can be also deducted from Table 2 that the image classification also suffers from imbalanced datasets and grabs the author's attention.

Table 2. The most popular author keywords used in publications

Words	Occurrences	Words	Occurrences	Words	Occurrences	Words	Occurrences
oversampling	877	feature selection	295	sampling	180	under-sampling	99
machine learning	791	imbalanced dataset	294	svm	178	parallel imaging	94
classification	701	data mining	276	convolutional neural network	147	convolutional neural networks	93
imbalanced data	672	random forest	265	big data	144	image reconstruction	93
smote	633	support vector machine	258	clustering	120	boosting	92
undersampling	502	imbalanced datasets	222	unbalanced data	120	neural networks	89
class imbalance	428	ensemble learning	219	decision tree	119	logistic regression	84
deep learning	423	imbalanced classification	184	over-sampling	109	adaboost	83
compressed sensing	334	imbalanced learning	184	support vector machines	99	anomaly detection	83

Fig. 7 demonstrates the most frequent words used over times based on the keywords plus that are terms or phrases that appear often in the titles of an article's references, but not necessarily in the title of the article or as author keywords [30]. The word cloud highlights smote, random forest, and undersampling. Imbalanced datasets in big data are also investigated by authors.

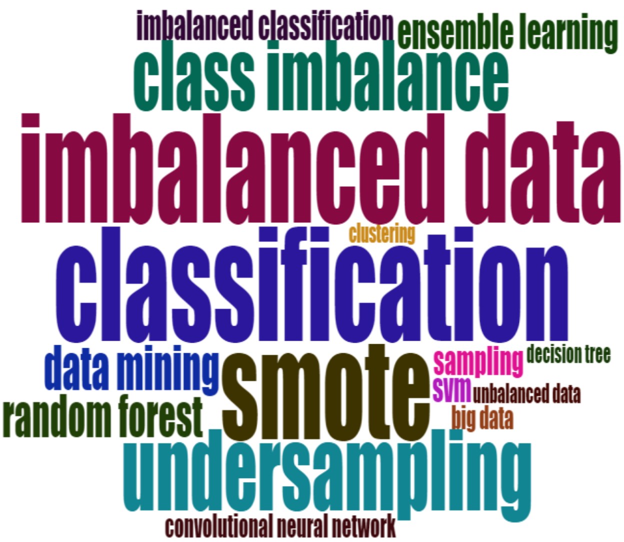

Figure 7. The frequency of the keywords used in imbalanced datasets

Fig. 8 shows the trend of the author's keywords over time that can help to choose better topics for future researches. According to the figure, oversampling, machine learning, classification, imbalanced data and SMOTE have a better growth compare to rest of the keywords.

3.6. Keywords Co-occurrence Network

Fig. 9 is keywords co-occurrence network (KCN) which depicts the connections between the author's keywords and puts forth knowledge structure in the field. As we can observe from the KCN, publications are grouped into five clusters. The orange and purple clusters are concerned with algorithmic solutions, whereas the red clusters are concerned with sampling techniques. The blue and green clusters are related to effects of the imbalanced datasets on more recent topics.

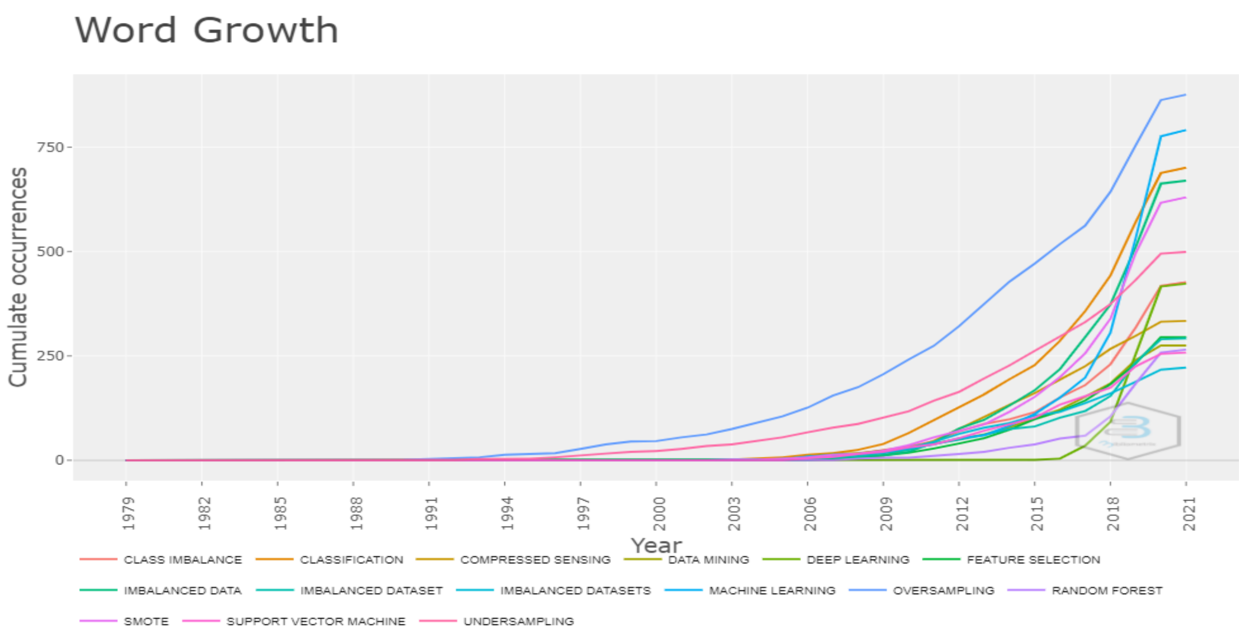

Figure 8. The cumulative word dynamics over time

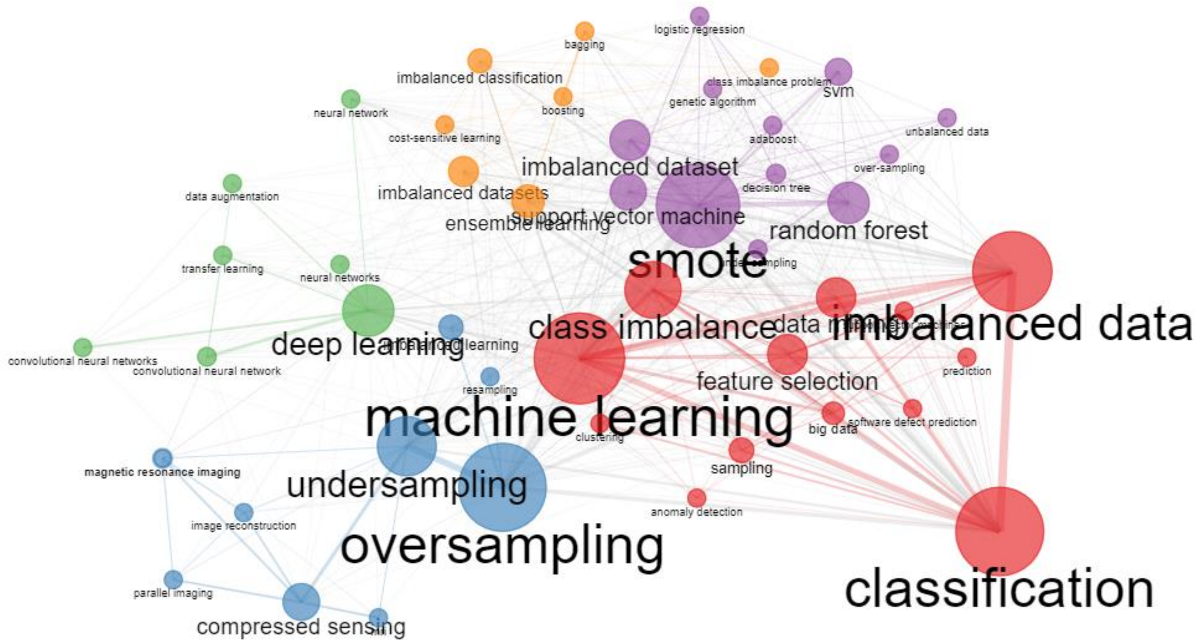

Figure 9. Keyword co-occurrence network

3.7. The Most Cited and the Most Contributed Authors

The analysis of the most cited authors provides us with an authors' list to follow fellow researchers in order to keep up in the area and stay updated with the latest developments. The frequency of most cited authors is shown in Table 3. Based on the Table 3, the most frequent author is Nitesh V. Chawla following Francisco Herrera and Lawrence O. Hall.

Table 3. The most cited authors

Author	Total Citations	Author	Total Citations
CHAWLA N V	5099	HE H	2546
HERRERA F	3957	KEGELMEYER W P	2437
HALL L O	3360	WANG Y	2428
BOWYER K W	3191	KHOSHGOFTAAR T M	1921
JAPKOWICZ N	2911	ZHANG Y	1893

Following active researchers in the field is as important as knowing the most cited authors to keep up with contemporary approaches. Fig. 10 shows the authors' productivity over the time. The size of the circles show the number of publications and the colors of the circles emphasize the total citations per year. According to Fig. 10, Taghi M. Khoshgoftaar and Yong Zhang are the most active researchers in the field.

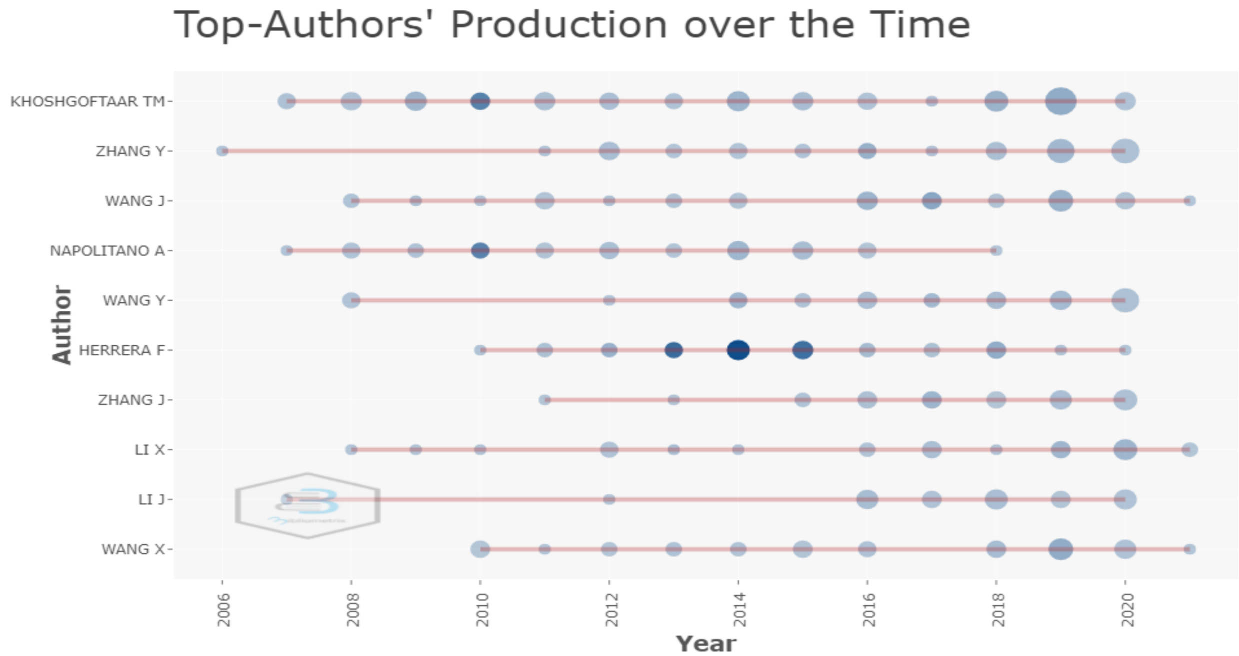

Figure 10. The authors' productivity over the time

E. Garfield [52] introduced the historiographic map as a graph to show a temporal network map of the most relevant direct citations from a bibliographic collection. The historiographic map is depicted in Fig. 11 that indicates the direction of progression in the field.

According to Fig. 11, a series of research started by Fernandez-Navarro et al. [53] in 2011 and a branch of research's series started by Lopez et al. [36] in 2013. Fernandez-Navarro et al. [53] proposed a dynamic over-sampling approach for imbalanced multiclass classification problem. Kovacs [54] published a comparative study of 85 variants of SMOTE in 2019. On top of Kovac's publication, Torres-Vasquez et al. [54] investigate the effect of balancing strategies on the Guillain–Barré Syndrome (GBS) dataset which is an immune system disorder. Lopez et al. [36] provide an overview of underlying problems of imbalanced dataset approaches to alleviate the effect of these problems. Rio et al. [55] analyzed the performance of the sampling strategies on big imbalanced datasets by using MapReduce. Yijing et al. [56] introduce an adaptive multiple classifier system for multiclass imbalanced dataset problem that includes three components (feature selection, resampling, ensemble learning). Kang et al. [57] introduce a new undersampling strategy to avoid noisy examples in minority class. Hasanin and Khoshgoftaar [58] researches effects of the random undersampling in the case of big imbalanced datasets with different degree of imbalanced ratio. Hasanin et al. [59] investigate imbalanced dataset problem in bioinformatics field and implement random undersampling with s feature selection approach. Abdel-Hamid et al. [60] highlighted the importance of border examples and propose a spark based mining framework that implements both oversampling and undersampling.

3.8. The Most Local Cited Sources

Choosing the ideal magazine to publish your work in may be a time-consuming procedure. Fig. 12 shows the most relevant sources in the field. According to the figure, Neurocomputing is the most cited source that describes current fundamental advances in the field of neurocomputing. The second most cited source is Expert Systems with Applications which is an international journal dedicated to the exchange of knowledge on expert and intelligent systems in business, government, and universities throughout the world and the following Machine Learning which publishes research papers on a wide range of learning approaches that have been applied to a variety of learning issues.

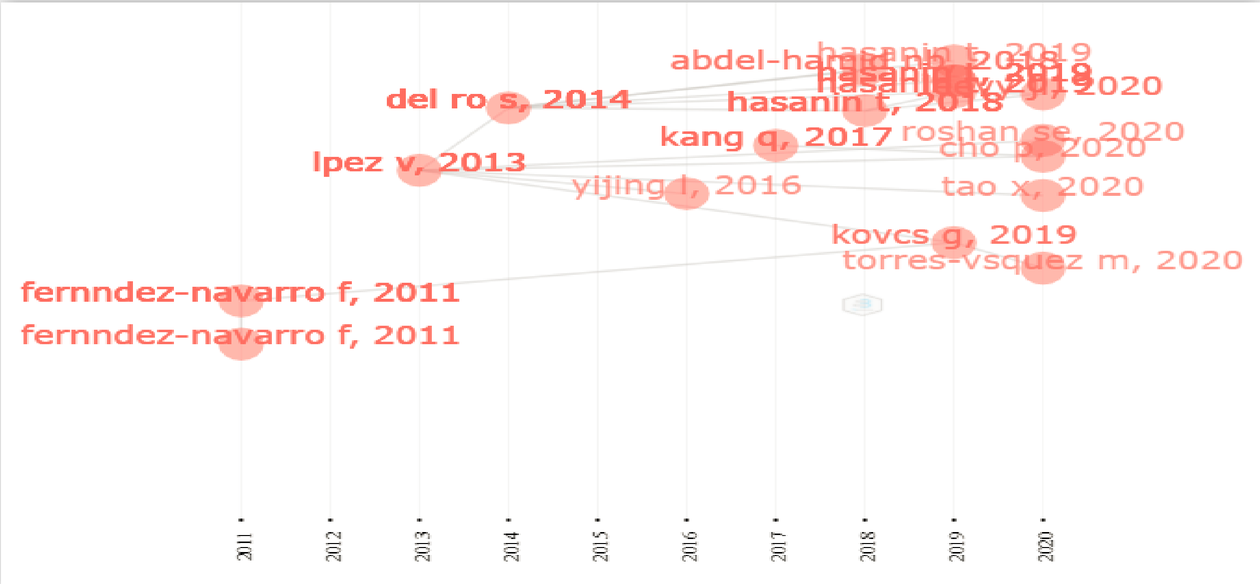

Figure 11. Historical direct citation network

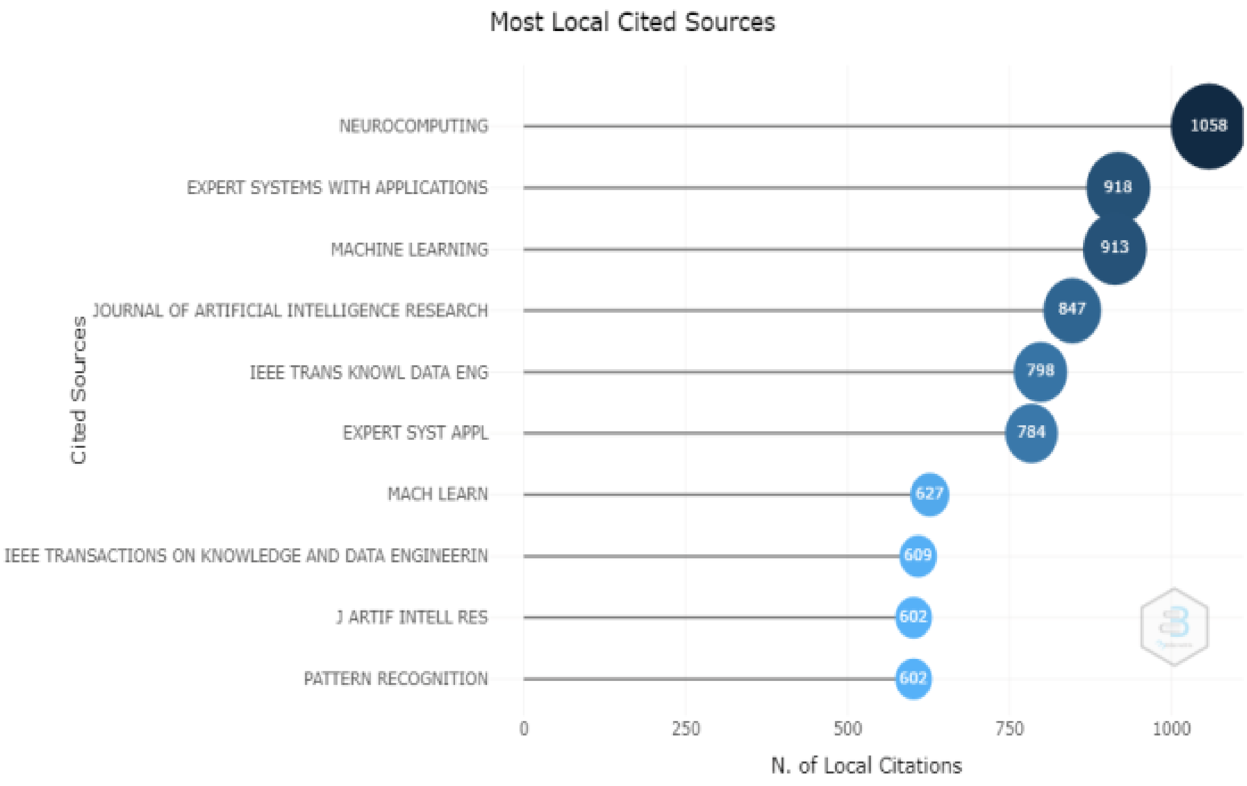

Figure 12. The most local cited sources

3.9. Thematic Map

The thematic map depicts a group of keywords and their relationships that are considered themes [61]. The thematic map has two dimensions are called density and centrality. The centrality lies in horizontal axis and shows the relations with other themes while the density lies in vertical axis and shows the within cluster connections. The thematic map can be analyzed based on its quadrants. According to thematic map in Fig 13, the important research clusters lie in the lower-right quadrant. The upper-left quadrant shows the specialized clusters. Motor themes are important and structure the field.

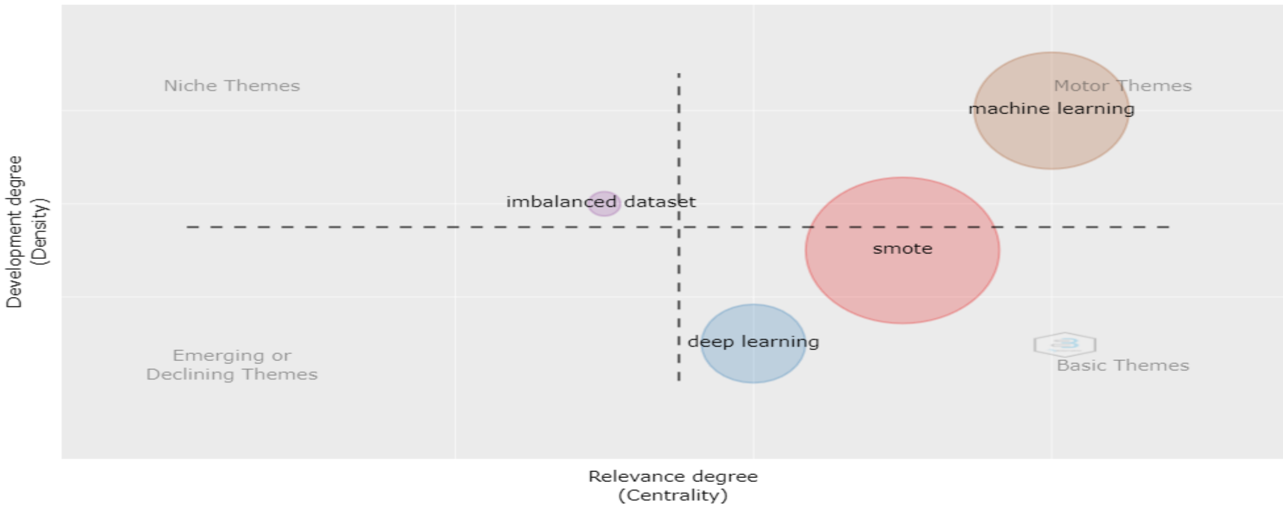

Figure 13. Thematic map

4. CONCLUSION AND FUTURE WORKS

The amount of data gathered and processed on a daily basis has also risen as a result of technological advancements. The massive volume of data posed its own set of challenges for typical machine learning models especially in the case of the imbalanced datasets. Imbalanced datasets, also known as unbalanced datasets, are considered as severely skewed distributions of target variables. This research aimed to offer a complete assessment of the literature related to the imbalanced datasets problem that has been published.

- The analysis based on Scopus showed that imbalanced datasets began to get attention in the early 2000s. Chawla et al. [62] published an editorial following the AAAI and ICML conferences and emphasize the importance of the problem. He and Garcia [18] also put these workshops among the major works in the field. After the workshops in regard of the problem and the publication of SMOTE increased the researches in the field.
- The imbalanced datasets problem is investigated mostly by authors from the United States, China, and Germany.
- SMOTE is still the most implemented solution despite the many years of research and numerous studies. SMOTE is mentioned by many authors in their papers [12, 22, 64, 65]. The interest in SMOTE also resulted in many variations of it. Safe-Level-SMOTE [43], SMOTE–IPF [49], Borderline-SMOTE [32] and various variants are introduced to improve the SMOTE performance. The keyword analysis also supports that oversampling and SMOTE grab a great attention of authors.

- Based on the keyword analysis the machine learning techniques random forest and SVM are mostly researched with imbalanced datasets. The researches on random forests and SVM model either focus on combining sampling techniques to improve the performance of imbalance datasets or algorithms are used to balance the datasets [37, 54, 63, 64].
- The most cited first authors in the field are Nitesh V. Chawla and Francisco Herrera. Chawla [31] especially contributes to the introduction of SMOTE whilst Herrera's reviews [64, 66] capture great attention. There is also the great contribution of the authors Taghi M. Khoshgoftaar and Yong Zhang in recent years. And based on the publications, the prominent journals in the area are Neurocomputing, Expert Systems with Applications, and Machine Learning.
- The researches in deep learning are accelerated by the amount of data. In recent years researchers started to work on the effects of imbalanced datasets on deep learning and big data models. The thematic map also shows that SMOTE and deep learning remain important. The imbalanced datasets on big data research are concentrating not only on obtaining a good

forecast, but also the speed of the proposed approaches [55, 58, 67].

- Sun et al. [11] publish the most thorough review of literature in the area. They emphasized that the researches in the field concentrated on binary classification which is can be also deducted from our bibliometric analysis results. There are still less ongoing studies on multiclass classification with imbalanced datasets and multilabel classification.
- Both Sun et al. [11] and our studies point out that ensemble approaches are caught interest to improve model performance.
- Our bibliometric analysis shows that imbalanced datasets in deep learning and big data have caught great attention with the recent developments and become more prevalent in the following years after the study [11].
- Our analysis shed a light on improvements in the field by focusing on quantitative results unlike the review by Sun et al. [11]

In this study, we focused on the analysis of metadata from SCOPUS. As an improvement to these results, the same study can be repeated by gathering data from the Web of Science. The more detailed researches can be also conducted based on the specific domain. Our results give a comprehensive analysis of the literature of all time. A limited examination of the literature, with a concentration on recent years, can be beneficial.

REFERENCES

[1] T. O. Ayodele, "Types of Machine Learning Algorithms", ***New Advances in Machine Learning***, 3, Yagang Zhang, Intech, Rijeka, Croatia, 19-48, 2010.

[2] G. E. Melo-Acosta, F. Duitama-Muñoz & J. D. Arias-Londoño, "Fraud Detection in Big Data Using Supervised and Semi-Supervised Learning Techniques", **IEEE Colombian Conference on Communications and Computing (COLCOM***)*, Cartagena, Colombia, 1-6, 2017.

[3] D. Zhang, H. Huang, Q. Chen, & Y. Jiang, "A Comparison Study of Credit Scoring Models", **Third International Conference on Natural Computation (ICNC 2007)**, Haikou, China, 1, 15-18, 2007.

[4] A. Maraş & S. Aydin, "Intercorrelation between Singular Spectrum of EEG Sub-Bands and Emotional States", ***National Conference on Electrical, Electronics and Biomedical Engineering (ELECO),*** Bursa, Turkey, 486-490, 2016.

[5] H. Asri, H. Mousannif, H. Al Moatassime & T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", *Procedia Computer Science*, 83, 1064-1069, 2016.

[6] V. A. Kumari & R. Chitra, "Classification of Diabetes Disease Using Support Vector Machine", International *Journal of Engineering Research and Applications*, 3(2), 1797-1801, 2013.

[7] J. Burez & D. Van den Poel, "Handling Class Imbalance in Customer Churn Prediction", Expert *Systems with Applications*, 36(3), 4626-4636, 2009.

[8] M. Anjaria & R. M. R. Guddeti, "A Novel Sentiment Analysis of Social Networks Using Supervised Learning", *Social Network Analysis and Mining*, 4(1), 181, 2014.

[9] R. Caruana & A. Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms", **Proceedings of the 23rd International Conference on Machine Learning**, Pittsburgh Pennsylvania, USA, 161-168, 2006.

[10] V. García, R. A. Mollineda, & J. S. Sánchez, "A New Performance Evaluation Method for Two-Class Imbalanced Problems", **Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)**, Orlando, FL, USA, 917-925, 2008.

[11] Y. Sun, A. K. Wong & M. S. Kamel, "Classification of Imbalanced Data: A Review", *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04), 687-719, 2009.

[12] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder & N. Seliya, "A Survey on Addressing High-Class Imbalance in Big Data", *Journal of Big Data*, 5(1), 1-30, 2018.

[13] Internet: Google (2021) Google Trends, http://www.google.com/trends/, 15.08.2021.

[14] F. Provost, "Machine Learning from Imbalanced Data Sets 101", **Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets**, Austin, Texas, USA, 68, 1-3, 2000.

[15] N. V. Chawla, N. Japkowicz & A. Koltcz, **Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets**, Washington, DC, USA, 2003.

[16] M. Zięba & J. M. Tomczak, "Boosted SVM with Active Learning Strategy for Imbalanced Data", *Soft Computing*, 19(12), 3357-3368, 2015.

[17] H. He & E. A. Garcia, "Learning from Imbalanced Data", *IEEE Transactions on knowledge and data engineering*, 21(9), 1263-1284, 2009.

[18] S. Wang, Z. Li, W. Chao & Q. Cao, "Applying Adaptive Over-Sampling Technique based on Data Density and Cost-Sensitive SVM to Imbalanced Learning", **The 2012 International Joint Conference on Neural Networks (IJCNN)**, Brisbane, QLD, Australia, 1-8, 2012.

[19] S. Belarouci & M. A. Chikh, "Medical Imbalanced Data Classification", *Advances in Science, Technology and Engineering Systems Journal*, 2(3), 116-124, 2017.

[20] Y. Yan, T. Yang, Y. Yang & J. Chen, "A Framework of Online Learning with Imbalanced Streaming Data", **Thirty-First AAAI Conference on Artificial Intelligence**, San Francisco, California USA, 2817-2823, 2017.

[21] A. Orriols-Puig & E. Bernadó-Mansilla, "Evolutionary rule-based Systems for Imbalanced Data Sets", *Soft Computing*, 13(3), 213, 2009.

[22] G. M. Weiss, "Mining with Rarity: a Unifying Framework", *ACM Sigkdd Explorations Newsletter*, 6(1), 7-19, 2004.

[23] T. Jo & N. Japkowicz, "Class imbalances versus small disjuncts". *ACM Sigkdd Explorations Newsletter*, 6(1), 40-49, 2004.

[24] R. C. Prati, G. E. Batista & M. C. Monard, "Class Imbalances versus Class Overlapping: An Analysis of a Learning System Behavior", **Mexican International Conference on Artificial Intelligence**, Mexico City, Mexico, 312-321, 2004.

[25] T. M. Khoshgoftaar, J. Van Hulse & A. Napolitano, "Comparing Boosting and Bagging Techniques with Noisy and Imbalanced Data", *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(3), 552-568, 2010.

[26] A. S. More & D. P. Rana, "Review of Random Forest Classification Techniques to Resolve Data Imbalance", **IEEE 1st International Conference on Intelligent Systems and Information Management (ICISIM)**, Aurangabad, India, 72-78, 2017.

[27] J. J. Ng & K. H. Chai, "A Bibliometric Analysis of Project Management Research", **IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)**, Singapore, 976-980, 2015.

[28] Internet: Scopus (2022) , https://www.scopus.com/, 26.05.2022.

[29] F. Machado & C. D. Martes, "Project Management Success: A Bibliometric Analisys", *Revista de Gestão e Projetos-GeP*, 6(1), 28-44, 2015.

[30] M. Aria & C. Cuccurullo, "Bibliometrix: An R-tool for Comprehensive Science Mapping Analysis", *Journal of Informetrics*, 11(4), 959-975, 2017.

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall & W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique", *Journal of Artificial Intelligence Research*, 16, 321-357, 2002.

[32] H. Han, W. Y. Wang & B. H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", **Advances in Intelligent Computing. ICIC. Lecture Notes in Computer Science**, Hefei, China, 878-887, 2005.

[33] N. V. Chawla, A. Lazarevic, L. O. Hall & K. W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting", **European Conference on Principles of Data Mining and Knowledge Discovery** , Cavtat-Dubrovnik, Croatia, 107-119, 2003.

[34] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince & F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-based Approaches", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews*), 42(4), 463-484, 2011.

[35] X. Y. Liu, J. Wu & Z. H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2), 539-550, 2008.

[36] V. López, A. Fernandez, S. Garcia, V. Palade, & F. Herrera, "An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics", *Information Sciences*, 250, 113-141, 2013.

[37] R. Akbani, S. Kwek & N. Japkowicz, "Applying support vector machines to imbalanced datasets", ***European Conference on Machine Learning,*** Pisa, Italy, 39-50, 2004.

[38] K. Veropoulos, C. Campbell & N. Cristianini, "Controlling the sensitivity of support vector machines", **Proceedings of the International Joint Conference on AI**, 55–60, 1999.

[39] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse & A. Napolitano, "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance", *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1), 185-197, 2009.

[40] Z. H. Zhou & X. Y. Liu, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem", *IEEE Transactions on Knowledge and Data Engineering*, 18(1), 63-77, 2005.

[41] A. Estabrooks, T. Jo & N. Japkowicz, "A Multiple Resampling Method for Learning from Imbalanced Data Sets", *Computational Intelligence*, 20(1), 18-36, 2004.

[42] Y. Tang, Y. Q. Zhang, N. V. Chawla & S. Krasser, "SVMs Modeling for Highly Imbalanced Classification", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1), 281-288, 2008.

[43] C. Bunkhumpornpat, K. Sinapiromsaran & C. Lursinsap, "Safe-Level-Smote: Safe-Level-Synthetic Minority Over-Sampling Technique for Handling the Class Imbalanced Problem", **Pacific-Asia Conference on Knowledge Discovery and Data Mining**, Bangkok, Thailand, 475-482, 2009.

[44] S. Wang, & X. Yao, "Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models", **IEEE Symposium on Computational Intelligence and Data Mining**, Nashville, TN, USA 324-331, 2009.

[45] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker & G. D. Tourassi, "Training Neural Network Classifiers for Medical Decision Making: The Effects of Imbalanced Datasets on Classification Performance", *Neural Networks*, 21(2-3), 427-436, 2008.

[46] M. Galar, A. Fernández, E. Barrenechea, & F. Herrera, "EUSBoost: Enhancing Ensembles for Highly Imbalanced Data-Sets by Evolutionary Undersampling", *Pattern Recognition*, 46(12), 3460-3471, 2013.

[47] A. Fernández, S. Garcia, F. Herrera & N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary", *Journal of Artificial Intelligence Research*, 61, 863-905, 2018.

[48] I. Triguero, M. Galar, S. Vluymans, C. Cornelis, H. Bustince, F. Herrera & Y. Saeys, "Evolutionary Undersampling for Imbalanced Big Data Classification", **IEEE Congress on Evolutionary Computation (CEC)**, Sendai, Japan, 715-722, 2015.

[49] J. A. Sáez, J. Luengo, J. Stefanowski & F. Herrera, "SMOTE–IPF: Addressing the Noisy and Borderline Examples Problem in Imbalanced Classification by a Re-sampling Method with Filtering", *Information Sciences*, 291, 184-203, 2015.

[50] L. Lusa, "Improved Shrunken Centroid Classifiers for High-Dimensional Class-Imbalanced Data", *BMC Bioinformatics*, 14(1), 1-13, 2013.

[51] I. Brown & C. Mues, "An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets", *Expert Systems with Applications*, 39(3), 3446-3453, 2012.

[52] E. Garfield, "Historiographic Mapping of Knowledge Domains Literature", *Journal of Information Science*, 30(2), 119-145, 2004.

[53] F. Fernández-Navarro, C. Hervás-Martínez, C. García-Alonso & M. Torres-Jiménez, "Determination of Relative Agrarian Technical Efficiency by a Dynamic Over-Sampling Procedure Guided by Minimum Sensitivity", *Expert Systems with Applications,* 38(10), 12483-12490, 2011.

[54] G. Kovács, "An Empirical Comparison and Evaluation of Minority Oversampling Techniques on a Large Number of Imbalanced Datasets", *Applied Soft Computing*, 83, 105662, 2019.

[55] S. Del Río, V. López, J. M. Benítez & F. Herrera, "On the Use of Mapreduce for Imbalanced Big Data Using Random Forest", *Information Sciences,* 285, 112-137, 2014.

[56] L. Yijing, G. Haixiang, L. Xiao, L. Yanan & L. Jinling, "Adapted Ensemble Classification Algorithm based on Multiple Classifier System and feature selection for classifying multi-class imbalanced data", *Knowledge-Based Systems,* 94, 88-104, 2016.

[57] Q. Kang, X. Chen, S. Li & M. Zhou, "A Noise-Filtered Under-sampling Scheme for Imbalanced Classification", *IEEE Transactions on Cybernetics,* 47(12), 4263-4274, 2016.

[58] T. Hasanin & T. Khoshgoftaar, "The Effects of Random Undersampling with Simulated Class Imbalance for Big Data", **IEEE International Conference on Information Reuse and Integration (IRI)**, Salt Lake City, UT, USA, 70-79, 2018.

[59] T. Hasanin, T. M. Khoshgoftaar, J. Leevy & N. Seliya, "Investigating Random Undersampling and Feature Selection on Bioinformatics Big Data", **IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)**, Newark, CA, USA, 346-356, 2019.

[60] N. B. Abdel-Hamid, S. ElGhamrawy, A. El Desouky & H. Arafat, "A Dynamic Spark-Based Classification Framework for Imbalanced Big Data", *Journal of Grid Computing*, 16(4), 607-626, 2018.

[61] H. Esfahani, K. Tavasoli & A. Jabbarzadeh, "Big Data and Social Media: A Scientometrics Analysis", *International Journal of Data and Network Science*, 3(3), 145-164, 2019.

[62] N. V. Chawla, N. Japkowicz & A. Kotcz, "Special Issue on Learning from Imbalanced Data Sets", *ACM SIGKDD Explorations Newsletter,* 6(1), 1-6, 2004.

[63] S. M. Abd Elrahman & A. Abraham, A Review of Class Imbalance Problem, *Journal of Network and Innovative Computing*, 1, 332-340, 2013.

[64] C. Su, S. Ju, Y. Liu & Z. Yu, "Improving Random Forest and Rotation Forest for Highly Imbalanced Datasets", *Intelligent Data Analysis*, 19(6), 1409-1432, 2015.

[65] F. Bulut, "Sınıflandırıcı topluluklarının dengesiz veri kümeleri üzerindeki performans analizleri", *Bilişim Teknolojileri Dergisi*, 9(2), 153, 2016.

[66] A. Fernández, S. García, & F. Herrera, "Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution", **International Conference on Hybrid Artificial Intelligence Systems**, Wroclaw, Poland, 1-10, 2011.

[67] D. J. Dittman, T. M. Khoshgoftaar & A. Napolitano,"The Effect of Data Sampling When Using Random Forest on Imbalanced Bioinformatics Data", **IEEE International Conference on Information Reuse and Integration**, San Francisco, CA, USA, 457-463, 2015.