Received: 1 July 2017 | Revised: 6 November 2017 | Accepted: 7 November 2017

DOI: 10.1002/widm.1239

WILEY WIREs DATA MINING AND KNOWLEDGE DISCOVERY

ADVANCED REVIEW

Evolutionary data mining and applications: A revision on the most cited papers from the last 10 years (2007–2017)

Rafael Alcalá[1] | María José Gacto[2] | Jesús Alcalá-Fdez[1]

[1]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

[2]Department of Computer Science, University of Jaén, Jaén, Spain

Correspondence
Jesús Alcalá-Fdez, Department of Computer Science and Artificial Intelligence, University of Granada, E-18071 Granada, Spain.
Email: jalcala@decsai.ugr.es

The ability of evolutionary algorithms (EAs) to manage a set of solutions, even attending multiple objectives, as well as their ability to optimize any kinds of values, allows them to fit very well some parts of the data-mining (DM) problems, whose native learning techniques usually associated with the inherent DM problem are not able to solve. Therefore, EAs are widely applied to complement or even replace the classical DM learning approaches. This application of EAs to the DM process is usually named evolutionary data mining (EDM). This contribution aims at showing a glimpse of the EDM field current state by focusing on the most cited papers published in the last 10 years. A descriptive analysis of the papers together with a bibliographic study is performed in order to differentiate past and current trends and to easily focus on significant further developments. Results show that, in the case of the most cited studied papers, the use of EAs on DM tasks is mainly focused on enhancing the classical learning techniques, thus completely replacing them only when it is directly motivated by the nature of problem. The bibliographic analysis is also showing that even though EAs were the main techniques used for EDM, the emergent evolutionary computation algorithms (swarm intelligence, etc.) are becoming nowadays the most cited and used ones. Based on all these facts, some potential further directions are also discussed.

This article is categorized under:
Fundamental Concepts of Data and Knowledge > Knowledge Representation
Technologies > Computational Intelligence
Technologies > Classification
Technologies > Prediction

1 | INTRODUCTION

Data mining (DM) (Han, Kamber, & Pei, 2011; Tan, Steinbach, & Kumar, 2006) can be described as the most important stage of the knowledge discovery of the data (KDD) process. It consists of the automatic process to discover interesting and unknown trends, patterns, and relationships on datasets, which otherwise would remain undetected. In other words, it tries to reveal the hidden information underlying large amounts of data. The wide range of DM techniques typically involve learning methods from the areas of machine learning (ML), statistics, and database systems, which depend on the type of DM problem being solved. For example, classification by neural networks (NNs) is usually solved by gradient descent network training, whereas decision trees are usually constructed by an iterative process that divides the data into subsets based on conditions that are set on the values of the problem attributes.

Even though evolutionary algorithms (EAs; Eiben & Smith, 2003) are not learning techniques, they have been also applied for learning and knowledge extraction. EAs are widely used optimization techniques that allow solving almost any combinatorial or continuous optimization problem, and even those involving both. These search techniques, which are population-based algorithms inspired on natural evolution and genetic processes, can be used as a complement to the standard DM learning approaches or even to replace them, since they can evolve descriptive or predictive models to their optimal

structure or parameters. This application of EAs to the DM process is nowadays an important part of what is widely known as evolutionary data mining (EDM).

Techniques that have been termed EAs have increased over time (Brabazon, O'Neill, & McGarraghy, 2015). During the 1960s and 1970s, evolution strategies (ESs), genetic algorithms (GAs), evolutionary programming (EP), and genetic programming (GP) were considered as the initial EAs. Lately in the 1980s and 1990s, learning classifier systems (LCSs) and differential evolution (DE) were also included as part of the EAs, considering this group of techniques as the origin of the so named evolutionary computation (EC) (Eiben & Smith, 2003) or evolutionary computing. Since they have been the most used in the literature, and therefore the most widely applied to DM, we will mainly focus on this set of techniques, belonging to the branch of EAs, in this contribution. However, nowadays there are many other types of evolutionary-inspired algorithms, that even though they do not fit with the EA's previous definition since they are not inspired on natural evolution and genetic processes, they are still based on populations or sets of solutions that cooperatively evolve toward a final optimum implementing intelligent behaviors, social interactions, etc. These more recent algorithms (with respect to the initial EAs) are nowadays considered together with the EAs as the EC (Yang, 2014) current family of algorithms. Since they truly represent a potential improvement to the EDM area, we will also pay attention to the "emergent" application on DM of these evolutionary techniques (Xing & Gao, 2014), namely emergent-EC algorithms from now on, by also analyzing their recent impact with respect to the application of the historically more used EAs.

The objective of this paper is to analyze the most cited and recent contributions to EDM for helping researchers to differentiate past and current trends in order to easily focus on significant further developments. To this end, we search for the five most cited proposals (EA-based) that focus on this field (five per year), from the last 10 years but also including 2017 to date, that is, 2007–2017, in the Computer Science category of the Clarivate Analytics ISI WoS.[1] Thus, we analyze two time windows, first five years (2007-2011) and last five years (2012-2017), as the medium/long-term past and the short-term past. We are considering the ISI WoS since even though it provides less inclusive search and indexing engines than other well-known resources, it is highly rigorous since it considers reliable citation sources in general, and for journals in particular. The most cited software and review papers are also briefly discussed as well as the most cited emergent-EC-based approaches. Moreover, we present a quick snapshot of the status of the publications on EDM (separately, EA and emergent-EC based) by analyzing number of papers and citations of the 10% most cited papers per year in the ISI WoS Computer Science area. We also draw visual science maps (Moya-Anegón et al., 2004) based on the free software Science Mapping Analysis Tool (SciMAT; Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2012) and The Open Graph Viz Platform (Gephi). Finally, we discuss the main current trends and possible further research directions.

Results show that, in the case of the most cited studied papers, the use of EAs on DM tasks is mainly focused on enhancing the classical DM techniques, thus completely replacing the typical classical DM algorithms only when it is directly motivated by the own problem nature. This is not always common in the lowest cited papers, which sometimes solve the whole DM problem using only EAs, thus forgetting there could be more suitable ML algorithms. The bibliographic analysis is also showing that even though EAs were the main techniques from the EC used for EDM, the emergent-EC algorithms (swarm, social intelligence, etc.) are becoming nowadays the most cited and used ones. Based on all these facts, some potential further directions are also discussed at the end of this paper.

This paper is organized as follows. Section 2 describes the search methodology we have considered for finding the most cited papers for each of the mentioned years. Section 3 analyzes the found proposals, locating them within the main areas of the DM and separating them into a first and a second five-year period. Section 4 focuses on works related to the available software tools and review papers on EDM, as well as the works applying the remaining new and recent techniques of the EC to the problem of DM. Section 5 presents a careful bibliographical study on EDM, revealing the main current research trends. In section 6, we discuss some critical considerations and possible further research directions. Finally, some concluding remarks are made in section 7.

2 | EDM MOST CITED PAPER SEARCH METHODOLOGY

This section overviews the methodology applied to search the most cited papers on EDM for each year, from 2007 to 2017, within the Computer Science category of the ISI WoS. In order to perform the search, we define the following two objectives:

- Analyzing the five most cited papers on EDM by year.
- Contrasting the most cited EDM papers with the emergent-EC-based DM papers.

TABLE 1 Terms used to perform the search at ISI WoS (* means 0 or more characters, $ 0 or 1)

(A) DM terms	(B) EA terms
data mining, machine learning,	differential evoluti*, evoluti* algorithm*,
data*driven, regression, predicti*, classif*,	evoluti* learning*, evoluti* multi*, evoluti* approach*,
unsupervised, semi*supervised, rule*	evoluti* strateg*, evoluti* tuning*, genetic algorithm*,
(C) Emergent-EC terms	evoluti* post$processing, evoluti* programming*,
swarm, colony, PSO, ACO, firefly, fruit fly,	genetic learning*, genetic multi*, genetic approach*,
gravitational evoluti*,	genetic tuning*, genetic programming,
estimation of distribution algorithm*,	genetic post$processing, gene expression programming,
EDA$, biogeography-based optimization,	learning classifier sytem*
imperialist competitive algorithm	**(D) Excluded terms**
	variable selection AND NOT (wrapper OR embedded),
	feature AND NOT (wrapper OR embedded),
	annealing, prototype selection, prototype learning,
	instance selection

DM = data mining; EA, evolutionary algorithm; EC, evolutionary computation.

In this sense, we perform two different searches, the second one including the first one, so that we can easily pick-up the most cited contributions and check for the differences between both search results:

1. Searching for the papers on EDM.
2. Searching for the papers on EDM but also including the emergent-EC-based approaches.

To do so, we make use of the terms showed in Table 1. These terms are grouped into four different categories: (A) terms related to DM; (B) terms related to EAs; (C) terms related to emergent-EC algorithms; (D) terms used to automatically exclude papers out of the searched topic. The final queries associated with the mentioned searches are

$$Q1: \text{Topic} = (A) \text{ AND } (B) \text{ AND NOT } (C) \text{ AND NOT Title} = (D),$$

$$Q2: \text{Topic} = (A) \text{ AND } (B) \text{ AND NOT } (C) \text{ AND NOT Title} = (D),$$

where "Topic" means that the terms are searched for in the Title, Abstract, Author Keywords, and Keywords Plus fields within each ISI paper record and where a category from Table 1 specification means searching for any of the terms in the category separated by comma in the table. The terms and the final queries were obtained by trial and error, trying to force discarding non-EDM papers but obtaining all the papers focused on EDM. As an example, directly using "evolutionary" or "evoluti*" as a term in (B) includes lots of papers coming from the biology area that even if they are devoted to DM they are out of the EDM scope. This is why this term appears many times in (B) but combined with a second term. Moreover, excluding papers where the title includes the terms in (D) help to exclude many papers devoted to the pre-processing stage of the KDD process (except papers using a wrapper or embedded approach, which involves learning) or several simulated annealing-based proposals.

Once the results are obtained, the table of percentiles from ESI WoK[2], period 2007–2017, is used in order to keep only the papers fitting the citation limits indicated in the table for the 10% most cited papers. Finally, a few papers out of the topic that were impossible to exclude by terminology or query modifications without eliminating some correct papers were carefully filtered by hand. From now on, we will name as Q1 and Q2 the final results from both queries once all the mentioned steps have been performed (query, 10%, hand filtering). These final records from Q1 and Q2 are the ones used to consider the second objective indicated at the beginning of this section, contrasting EA-based with emergent-EC-based DM, as well as, the five most cited papers are taken from the Q1 results by year. These queries were performed in June 16, 2017.

3 | MOST CITED PAPERS ON EDM (2007–2017)

In this section, we present and analyze the five most cited papers per year in the 2007–2017 period. As mentioned, we have applied the Q1 query to obtain the contributions on EDM by year, then selecting the five top cited ones. In Figure 1, we summarize the found proposals by considering a division by the topics of the DM area. We try to show how many proposals

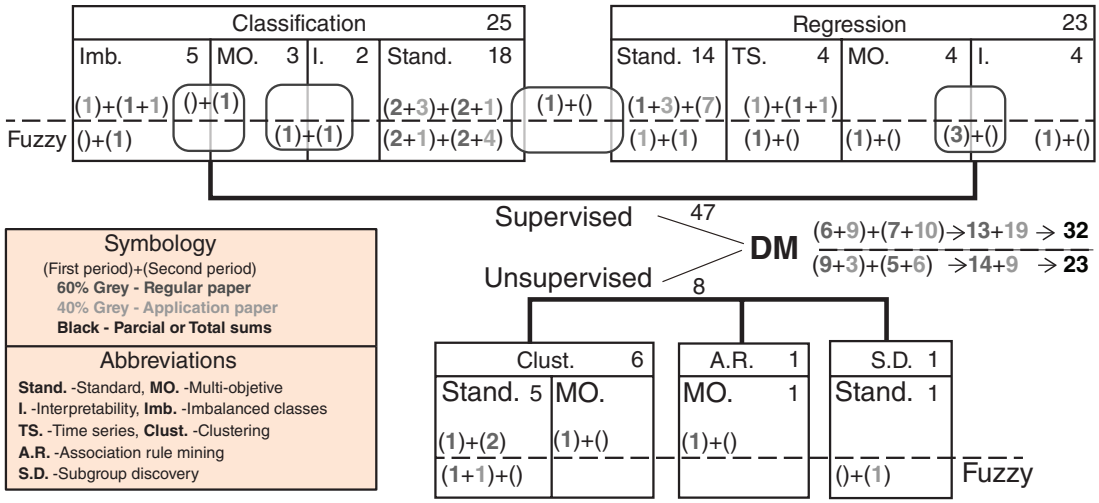

FIGURE 1 Summary of the five most cited papers per year in the periods 2007–2011 (first period) and 2012–2017 (second period) by data-mining problem type

are within each category, how many proposals are from the first five-year period (2007–2011) or from the last 6-year period (2012–2017), as well as how many proposals are representing application papers. Since there are many fuzzy logic-based proposals within the most cited contributions in the particular case of the EDM, we have also reflected it in the figure. At this point, we would like to remark that the papers counted within the interpretability category are those including specific mechanisms devoted to obtain more interpretable models, that is, they are not only including the word "interpretability" in the title or abstract but enhancing or handling the interpretability as an important part of the proposed algorithm.

In order to do so, we are distinguishing between numbers of papers in the first and the second periods by including this information into consecutive brackets, where numbers in the first bracket refer to the first period and numbers in the second bracket refer to the second one, (first) + (second). Empty brackets mean there are no proposals on a topic on the corresponding period. A square with round corners spreading across two boxes in the figure means that the papers here belong to the topics of both boxes. This is why the sum of the number of papers in some sub-topics is not exactly the same at the main topic (since papers counted into two sub-topics must be counted once for a topic). The dashed line is separating the counts for non-fuzzy papers (up) from the counts for fuzzy papers (down). Real-world application papers are also distinguished from standard papers by using a different color from the grey scale. Finally, a summary considering supervised and unsupervised proposals together is shown after the letters DM by periods and after the arrows by joining both periods (2007–2017).

Observing the numeric values in the figure and always taking into account that we are only talking about what is appearing on the top most cited papers, we can say that most of the proposals are in the fields of classification and regression, 25 and 23, respectively, while a few ones, only 8, are devoted to unsupervised learning (whose majority are for clustering). This is also notable that a 50.9% are specific real-world application papers, which is even higher in the second period, thus showing a tendency on the increasing importance of these kinds of applied contributions. Moreover, fuzzy logic-based proposals constitute 41.81% of the total number of contributions.

In the following, the papers found for the first five-year period are presented and briefly analyzed. Then, the second period is also discussed. Focusing on the distribution in the different fields shown in Figure 1, we will subdivide each period grouping the proposals into classification, regression, and unsupervised approaches.

3.1 | First period (2007–2011)

Table 2 includes information about the main characteristics of the studied methods as well as their citations for the first period (Q1 query). Papers that are classified as "Highly cited papers" by Clarivate Analytics in the Essential Science Indicators (they are within the top 1% most cited papers for any of the ISI categories considering all the papers, not only EDM papers, published in the same year) have been also marked in the table. It also includes information on the problem being solved (Application), where a dash means standard classification, regression, etc. The DM task being solved is also included in the table (DM problem), as well as the Model and EA types (DM part and evolutionary part). Finally, this table shows the techniques used for the proposal (Algorithm) and the validation methodology used (Val.). In the case of the Algorithm column, & means that both techniques are performed together within the same process and + means that both techniques are performed separately in two different stages.

TABLE 2 The five EDM most cited papers in Computer Science from 2007 to 2011

Year	References	Cites	Application	DM problem	Model type	EA type	Algorithm	Val.
2007	Handl and Knowles (2007)[a]	250	–	Unsup./ MO.Clu.	Cluster	MOEA	Clusters & PESA-II	HO & ST
	Huang, Chen, and Wang (2007)[b]	213	Economy	Clas.	SVM	GA	GA (feat.) + SVM & GA	CV
	Ishibuchi and Nojima (2007)	201	–	Clas./I.	FRBCS	MOEA	FRBS & NSGA-II	CV
	Wu, Tzeng, Goo, & Fang (2007)[b]	165	Economy	Reg./TS	SVM	GA	SVM & GA	HO & Bo. & ST
	Zhu, Ong, and Dash (2007)	151	–	Clas.	KNN	MMs + 1NN	1-NN & GA (wrapper)	HO
2008	Ding and Zhang (2008)[b,c]	131	Biology	Clas./Ens.	F KNN Ens.	Immune GA	(F KNNs) & IGA (Ens.)	HO
	Baykasoglu, Guellue, Canakci, and Oebakir (2008)	95	Geology	Reg.	SR	GP	MEP, GEP, LGP	HO & ST
	Mansoori, Zolghadri and Katebi (2008)	79	–	Clas.	FRBCS	GA (steady-state)	FRBS & SSGA$^{(SGERD)}$	CV
	Alatas, Akin, and Karci (2008)	78	–	Unsup./ARs	AR	MO DE	ARs. & MO-DE$^{(MODENAR)}$	HO
	Rojas et al. (2008)	73	–	Reg./TS	SR + F Rules + NN	GA	ARMA + F Rules + NN & GA	HO
2009	Garcia, Fernandez, Luengo, and Herrera (2009)[a,c]	254	–	Clas./ Comparison	Many	Many	GBML approaches	CV & ST
	Martinez, Castillo, and Aguilar (2009)[a]	166	Mobile robot	Reg.	T2 TSK FRBS	GA	FRBS & GA	HO
	Suresh, Babu, and Kim (2009)	110	Image	Clas./Imb.MC.	NN	GA	ELM(NN) & GA$^{(RCGA-ELM)}$	HO
	Alcala, Ducange, Herrera, Lazzerini, and Marcelloni (2009)	90	–	Reg./I.	FRBS	MOEA	FRBS & PAES	CV & ST
	Fei and Zhang (2009	81	Energy	Clas.	SVM	GA	SVM & GA	HO
2010	Jose Gacto, Alcala, and Herrera (2010)	80	–	Reg./I.	FRBS	MOEA	FRBS & SPEA2$_{ACC}$ (TS$_{SP2\text{-}SI}$)	CV & ST
	Cheng, Chen, and Wei (2010)	74	Stock Market	Reg.	SR	GA	Rought Sets + GA	HO
	Lei, Zuo, He, and Zi (2010)	69	Engineering	Clas./Ens.	NN & KNN	GA	(MLP, RBF, KNN) & GA (Ens.)	HO
	Das and Sil (2010)	67	Image	Unsup./Clu.	F Clusters	DE	F clusters & DE	HO
	Pulkkinen and Koivisto (2010)	67	–	Reg./I.	FRBS	MOEA	(C4.5 + WM) & NSGA-II	CV & ST
2011	Gandomi and Alavi (2011)[a]	98	–	Reg.	SR	Multi-Stage GP	MSGP	HO
	Alcala-Fdez, Alcala, and Herrera (2011)[a]	93	–	Clas./ARs/HD	FRBCS	GA	A priori & CHC$^{(FARC-HD)}$	CV & ST
	Ghosh, Mishra, and Ghosh (2011)	92	Image	Unsup./Clu.	F Clusters	GA	FCM & GA, GKC & GA	HO
	Alavi and Gandomi (2011)[b]	82	Engineering	Reg.	NN & SR	GP	MEP, GEP and LGP	HO
	Alcala, Jose Gacto, and Herrera (2011)	70	–	Reg./HD/Sc.	FRBS	MOEA	FRBS & SPEA2$_{E/E}$	CV & ST

[a] Highly cited paper in computer science

[b] Highly cited paper in engineering.

[c] Available software.

ARs = association rules; Bo. = boosting; Clas. = classification; clus. = clustering; CV = cross fold validation; DE = differential evolution; Ens. = Ensemble; F = fuzzy; FCM = fuzzy C-means; FRBS = fuzzy rule-based system; GKC = Gustafson–Kessel clustering; FRBCS = fuzzy rule-based classification system; GEP = gene expression programming; HD = high dimensionality; HO = hold-out; I. = interpretability; Imb. = imbalanced; LGP = linear genetic programming; MC. = multi-class; MEP = multi-expression programming; MM = memetic; MO = multi-objective; Reg. = regression; Sc. = scalability; SR. = symbolic regression; STs = statistical tests; TS = time series; T2 = type-2; Unsup. = unsupervised.

We can observe that there is a wide variety of particular applications related to economy, biology, geology, etc., among the most cited papers, which are more or less equally distributed through the years. This shows the interest these applied proposals generate and their relevance to the research community. However, it is also remarkable that most of them are applying standard versions of GAs as well as the simplest validation model, the hold-out (HO) method, which is not a statistically robust evaluation method. Another remarkable fact is that only 9 of the 25 references in Table 2 are using statistical tests (STs). These facts were not expected in papers that accumulate so many citations.

3.1.1 | Classification (first period)

From the characteristics shown in the table, there are different DM techniques applied to classification tasks. Out of the nine references addressing classification in Table 2, as observed in the column Model type, four references used some kind of fuzzy system: three using fuzzy rule-based classification systems (FRBCSs) and one using fuzzy *k*-nearest neighbors (Ding & Zhang, 2008) (*k*-NN) ensembles.

The two support vector machines (SMVs) for classification are only used in application papers. These two papers apply GAs to learn the SVM input parameters in order to get the SVM algorithm learning better (Huang et al., 2007; Fei & Zhang, 2009). We can also find ensembles using *k*-NN and neural networks (Lei et al., 2010) (NNs), where EAs are applied for finding the best subset of models' combination (selection of models) or learning the weights for combining all the individual models. fuzzy rule-based systems (FRBSs) are also used for classification, where the EAs are used for learning the whole system (Ishibuchi & Nojima, 2007; Mansoori et al., 2008) or as post-processing (Alcala-Fdez, Fernandez, et al., 2011).

We can also find two particular contributions. The use of NNs for imbalanced multi-class classification (Suresh et al., 2009) and a proposal for the use of statistical techniques in the analysis of the behavior of genetic-Based machine learning algorithms(Garcia et al., 2009).

3.1.2 | Regression (first period)

There is only a SVM-based approach applied to time series (TS) forecasting (CH et al., 2007) whose combination with a GA is quite similar to the mentioned approaches for classification. FRBSs are mainly combined with multi-objective EAs (MOEAs) in order to take into account interpretability issues (Alcala et al., 2009; Jose Gacto et al., 2010; Pulkkinen & Koivisto, 2010) or to decrease the search space on high-dimensional problems (Alcala et al., 2011). There is also a particular case making use of type-2 fuzzy logic together with a GA in order to control a robot (Martinez et al., 2009). Symbolic regression is also accomplished by means of gene expression programming (GEP) or GP (Alavi & Gandomi, 2011; Baykasoglu et al., 2008; Gandomi & Alavi, 2011), but also by a GA (Cheng et al., 2010) or in combination with the derivation of fuzzy rules (Rojas et al., 2008),

3.1.3 | Unsupervised (first period)

There are only two approaches devoted to clustering: One of them performed multi-objective clustering (Handl & Knowles, 2007) by means of a known MOEA and the other performed fuzzy clustering by improving the Fuzzy C-Means algorithm in a specific image clustering problem (Ghosh et al., 2011). And finally multi-objective association rule (AR) mining (Alatas et al., 2008) is also performed by a multi-objective DE.

3.2 | Second period (2012–2017)

Table 3 includes information of the main characteristics of the studied methods as well as their citations for the second period (Q1 query). A description of this kind of table can be found in the previous subsection ("First period [2007–2011]"). Since there are a few more papers with one citation than the five ones shown in the table for 2017 (papers with 0 citations were removed), we have used U1, usage count (last 180 days), provided by ISI WoS as tie-breaking criterion to select the ones included in the table. Anyway, we have to point out that these papers should be considered only as representative examples, since they will surely change their position in a near future.

As we can see, there are even more specific application papers. The use of standard versions of GAs is still a frequent issue, but in these last years there are also some approaches using more advanced algorithms such as the CHC algorithm, the cooperative coevolution (CC) and a hybrid combination of ESs and GAs.

There are also more proposals using STs and cross-validation. However, there are still only 12 papers (out of 25) using STs and 15 papers are still using the simple (but not statistically robust) HO method as the only evaluation method.

3.2.1 | Classification (second period)

From the characteristics shown in Table 3, the following approaches are applying to classification tasks. Two SVM approaches appear again as for the previous period, where EAs are performing the learning of the SVM algorithm's input parameters. One of them is directly applying GAs,(Kuang et al., 2014) while the other is including a new twin SVM proposal tuned by a GA (Shao et al., 2013). The first one was devoted to solve specific real-world applications, while the second one is the only SVM approach in all the study (2007–2017) that is not an application paper but a proposal for standard classification.

NNs have been also obtained by application of the DE for classification (Cao et al., 2012) (but also for regression in this case), and by hybrid application of Grammatical Evolution (topology) and GAs (weights) (Ahmadizar et al., 2015).

In Table 3, out of the 16 articles addressing a classification problem, 7 use some kind of fuzzy system: 6 using FRBCS, and 1 using fuzzy K-NN. More precisely, FRBSs are again obtained by means of GAs, with CHC applied to tune interval-valued fuzzy systems (Antonio Sanz et al., 2013, 2014) which is lately applied to imbalanced classification (Antonio Sanz et al., 2015), or with a GA applied in a multi-stage method for rule optimization (Nguyen et al., 2015). Interpretability of FRBSs is also taken into account (Rudzinski, 2016) by means of MOEAs this past year. Interval-valued fuzzy kNN have

TABLE 3 The five EDM most cited papers in Computer Science from 2012 to 2017

Year	References	Cites	Application	DM problem	Model type	EA type	Algorithm	Val.
2012	Gandomi and Alavi (2012a)[a]	71	Engineering	Reg.	SR	GP (MG GP)	MGGP	HO
	Chandra and Zhang (2012)	63	–	Reg./TS	NN	CC GA	Elman-RNN & CC G3-PCX	HO
	Huang (2012)	55	Stock Market	Reg.	SVM	GA	SVR & GA	HO
	Cao, Lin, and Huang (2012)	54	–	Clas.&Reg.	NN	DE	ELM(NN) & Self-adaptive DE$^{(SaE\text{-}ELM)}$	HO
	Gandomi and Alavi (2012b)	45	Engineering	Reg.	SR	GP (MG GP)	MGGP	HO
2013	Kisi, Shiri, and Tombul (2013)	42	Geology	Reg./TS	SR	GP	SR GEP	HO
	Antonio Sanz, Fernandez, Bustince, and Herrera (2013)[b]	39	–	Clas.	IV FRBCS	GA	A priori & CHC $^{(IVTURS-FARC)}$	CV & ST
	Shao, Wang, Chen, and Deng (2013)	38	–	Clas.	SVM	GA	Twin SVM & GA	CV
	Cpalka, Rebrova, Nowicki, and Rutkowski (2013)	36	–	Reg.	NFS	ES(μ, λ)	NFS & ES (μ, λ)	HO
	Bhowan, Johnston, Zhang, and Yao (2013)	36	–	Clas./Imb.	SR	MO GP	MO-GP (Ens.)	HO & ST
2014	Cpalka, Lapa, Przybyl, and Zalasinski (2014)	39	–	Reg./I.	NFS	ES(μ, λ) and GA	NFS & ES (μ, λ)-GA	HO
	Krawczyk, Wozniak, and Schaefer (2014)	39	–	Clas./Imb.	DT	GA	Cost-Sensitive DT + GA	CV & ST
	Kuang, Xu, and Zhang (2014)	38	Security	Clas.	SVM	GA	Kernel-PCA + SVM & GA	HO
	Antonio Sanz et al. (2014)	36	Medicine	Clas.	IV FRBCS	GA	FRBS & CC GA + CHC	CV & ST
	Menendez, Barrero, and Camacho (2014)	35	–	Unsup./ C Clu.	Cluster	GA	CSG & GA $^{(GGC)}$	HO
2015	Chandwani, Agrawal, and Nagar (2015)	25	Engineering	Reg.	NN	GA	ANN & GA	HO
	Ahmadizar, Soltanian, AkhlaghianTab, and Tsoulos (2015)	16	–	Clas.	NN	GE + GA	ANN & GE (topology) & GA (weights)	CV & ST
	Nguyen, Khosravi, Creighton, and Nahavandi (2015)	16	Medicine	Clas.	FRBCS	GA	FRBS & AVQC (RI) + GA (RO) + GDSL (PT)	CV & ST
	Carmona et al. (2015)	15	Medicine	SD	FRBS	GP	GP	CV & ST
	Antonio Sanz, Bernardo, Herrera, Bustince, and Hagras (2015)	14	Economy	Clas./Imb.	IV FRBCS	GA	A priori & CHC $^{(IVTURS-FARC)}$	CV & ST
2016	Wang, Wang, and Liu (2016)[a]	11	–	Clas./MV TS	NN	DE	RNN + adaptive DE (parameters)	CV & ST
	Gorzalczany and Rudzinski (2016)	8	Economy	Clas.	FRBCS	MOEA	FRBS & NSGA-II	CV
	Derrac, Chiclana, Garcia, and Herrera (2016)	8	–	Clas.	IV F-KNN	GA	fuzzy-kNN + CHC	CV & ST
	Rudzinski (2016)	8	–	Clas./I.	FRBCS	MOEA	FRBS & (SPEA2, NSGA-II)	CV & ST
	Krawczyk, Galar, Jelen, and Herrera (2016)	7	Medicine	Clas./Imb., Ens.& Bo.	KNN	EU	(FCM, LS, GLS) + C4.5 s & CHC 1NN EU (Ens.)$^{(EUSBoost)}$	Bo. & CV & ST
2017	Oliveira et al. (2017)	1	–	Unsup./Clu./BD	Cluster	GA	k-means & GA (MR), k-means & GA (Ens. MR)	HO & ST
	Duchanoy et al. (2017)	1	Engineering	Reg.	NN	DE	RNN & DE	HO
	Shen et al. (2017)	1	Medicine	Reg./BD	SVM	GA	SVM & GA (MR)	HO
	Demertzis, Iliadis, Avramidis, and El-Kassaby (2017)	1	Geology	Reg.	NN & SR	GP (GEP)	Feed-Forward ANN, GEP SR	HO
	Serdio et al. (2017)	1	Engineering	Clas.	SFIS	GA (MM)	SFIS & GA T (Emb.)$^{(GenSparse\text{-}FIS)}$	HO

[a] Highly cited paper in engineering.

[b] Available software.

AVQC = adaptive vector quantization clustering; BD = big data; Bo. = boosting; C = continuity; CC = cooperative coevolution; Clas. = classification; Clus. = clustering; CSG = cluster similarity graph; CV = cross fold validation; DE = differential evolution; DT = decision trees; Emb. = embedded; Ens. = ensemble; ES = evolutionary strategy; EU = evolutionary undersampling; FCM = fuzzy C-means; FRBS = fuzzy rule-based system; FRBCS = fuzzy rule-based classification system; GDSL = gradient descent supervised learning; GE = grammatical evolution; GEP = gene expression programming; GLS = grey-level segmentation; HO = holdout; I. = interpretability; Imb. = imbalanced; IV = interval-valued; LS = level-set; MG = multi-gene; MM = memetic; MO = multi-objective; MR = MapReduce; MV = multivariate; NFS = neuro-fuzzy system; PT = parameter tuning; Reg. = regression; RI = rule initialization; RO = rule optimization; SD = subgroup discovery; SR = symbolic regression; SFIS = sparse-fuzzy inference system; ST = statistical tests; T = tuning; TS = time series; Unsup. = unsupervised.

been also obtained by means of CHC optimization (Derrac et al., 2016), as well as sparse fuzzy inference systems were obtained by the application of a memetic GA (Serdio et al., 2017).

Multi-objective GP is also used to perform symbolic regression for imbalanced classification (Bhowan et al., 2013). Imbalanced classification is also addressed (Krawczyk et al., 2014) by learning cost-sensitive decision trees applying a GA or proposing evolutionary undersampling by the CHC application in order to obtain 1NN-based ensembles (Krawczyk et al., 2016).

Finally, a particular interesting contribution is the one solving multi-variate TS classification. This is a relatively recent problem addressing the interactions among many uni-variate TS. In this case, the parameters of a recurrent NN are evolved by means of an adaptive DE (Wang et al., 2016).

3.2.2 | Regression (second period)

In Table 3, out of the 11 articles that are addressing a regression problem, 5 articles use NNs as a model type, and 4 articles use GP for symbolic regression.

An ES is applied to learn a neuro-fuzzy system (Cpalka et al., 2013), then lately including interpretability issues (Cpalka et al., 2014) and the use of a hybrid ES-GA.

GP is again applied to perform symbolic regression (Gandomi & Alavi, 2012a), including TS forecasting (Kisi et al., 2013). Multi-gene GP can be also found to perform symbolic regression (Gandomi & Alavi, 2012b) as well as GEP (Demertzis et al., 2017).

Artificial NNs are also constructed by the application of a GA (Chandwani et al., 2015), including recurrent NNs for TS forecasting by means of a cooperative coevolutionary GA (Chandra & Zhang, 2012). Besides, a DE also evolves recurrent NNs for standard regression (Duchanoy et al., 2017).

As for the previous period, a GA is used to learn the optimal input parameters of an SVM (Huang, 2012) for a specific real-world application, including their use within a MapReduce (MR) framework for solving medical big data problems (Shen et al., 2017).

3.2.3 | Unsupervised (second period)

In the case of the EDM application to unsupervised problems in this second period, a GA is applied for continuity clustering by learning cluster similarity graphs (Menendez et al., 2014). Subgroup discovery (SD) is also addressed in this period by means of GP applied to the derivation of fuzzy rules (Carmona et al., 2015) in a medicine-specific application. Finally, clustering is performed on big data problems by means of *k*-means-based ensembles optimized by a GA within an MR framework (Oliveira et al., 2017).

3.3 | Applications in both periods (2007–2011 and 2012–2017)

The ability of EAs to manage a set of solutions, even attending to multiple objectives, as well as their ability to optimize any kinds of values, allows them to be successfully applied in a wide variety of applications. Tables 2 and 3 shows a list of the specific problems that have been solved by EAs on the top most cited papers in the periods 2007–2011 and 2012–2017 (see column "Application"), respectively. In this list, we can find a great variety of problems that are related to very diverse subjects. At a glance, we can find applications in economy, biology, geology, energy, image, mobile robot, engineering, stock market and security, highlighting the number of proposals for applications in the areas of image (Eiben & Smith, 2003), economy (Han et al., 2011), and engineering (Han et al., 2011) in the first period, and for the engineering (Yang, 2014) and medicine (Yang, 2014) in the second. This demonstrates the great capacity of these methods to be applied to a great variety of problems.

Analyzing the number of contributions in these periods, we can see how more than 44% of papers in the first period (2007–2011) are specific real-world application papers and the ratio is even higher in the second period (2012–2017), reaching more than 57% of studied papers. There is also an evidence of the interest that is awakening in the last years these methods to solve real-world problems. Moreover, more than 40% of the application papers studied are fuzzy logic-based proposals due to their ability to better address the imprecision and uncertainty and to incorporate expert knowledge and granular computing (Yao, Vasilakos, & Pedrycz, 2013).

4 | SOFTWARE TOOLS, REVIEW PAPERS, AND REMAINING EC TECHNIQUES ON DM

Here, we include interesting papers that, because they are not algorithm-based proposals or because they are not using EAs, got out of the previous analysis. They are the papers devoted to software tools including EDM algorithms, review papers highly related to EDM and some interesting application examples of the emergent-EC algorithms to DM.

4.1 | Software papers

Since each year, EDM proposals are being developed more and more, there is also a growing interest in the development of software tools that compile all these proposals according to a threefold necessity: (a) enabling researchers with low-medium knowledge to apply these proposals to their problems successfully; (b) to compare the authors' proposals versus other proposals of the literature on the topic; and (c) to carry out a complete experimentation to select the best suited solution for our problem.

Table 4 shows the most referenced software tools in the last years. Only 4 publications were found in the complete period (2007–2017) from the Q1 query.

The KEEL (Knowledge Extraction based on Evolutionary Learning) software suit (Alcala-Fdez et al., 2009, 2011; Triguero et al., 2017) allows researchers to evaluate the performance of evolutionary learning for different kinds of DM problems: regression, classification, clustering, and so on. Shark (Igel et al., 2008) is a library that provides single and multiobjective methods for regression and classification tasks. Finally, DREAM (Vrugt, 2016) is a toolbox that allows us to apply the DE Adaptive Metropolis algorithm (DREAM) to regression problems.

4.2 | Review papers

Review papers are often among the most cited contributions because of their usefulness in positioning future articles in a specific theme. We have preferred to leave them apart from the study of the most cited contributions since they do not include any new specific proposal. However, as they are what could be considered as the base bibliography to any bibliographic study, we also introduce them in this section. In this way, while we are focusing on the most cited and recent papers in order to see the most interesting current trends, these reviews usually include a significant part of the remaining proposals and/or their categorization.

Table 5 lists the review works found within the 2007–2017 period. Only 7 publications were found in the complete period (2007–2017) from the Q1 query.

In the following, we indicate the specific topic they cover. In 2010 (Fernandez et al., 2010), a state-of-the-art summary and taxonomy is provided for the genetic-based machine learning algorithms for rule induction in classification tasks. Foundations, algorithms, and applications of SD are reviewed in 2011 (Herrera et al., 2011). The authors provided in 2012 (Barros et al., 2012) a detailed survey of EAs to evolve decision trees for classification and regression. In 2013 (Fazzolari et al., 2013), an overview of multiobjective evolutionary fuzzy systems, describing the main contributions on this field and providing a two-level taxonomy is performed. Part I and Part II of a review (Mukhopadhyay et al., 2014a; Mukhopadhyay et al., 2014b) on multiobjective EAs for DM is also proposed in 2014. In 2015, the authors review (Fernandez et al., 2015) the progression of the so named evolutionary fuzzy systems by analyzing their taxonomy and components. Finally, a revision (Alcala-Fdez & Alonso, 2016) on the existent freely available and open-source fuzzy systems software is performed in 2016, which includes EAs application to FRBSs, neuro-fuzzy systems, and fuzzy AR mining.

4.3 | Application of the emergent-EC techniques

The following section will show as the emergent-EC-based publications that appear within the EDM most cited papers in the past 10 years grow in quantity and importance across the years. Here, we only introduce some representative examples of the great variety of emergent-EC algorithms that have been applied to the DM problems (see Reference (Xing & Gao, 2014) where a detailed description of 134 of these emergent-EC algorithms can be found). For these representatives, we will only consider those appearing in the top five contributions obtained from the Q2 query.

TABLE 4 Existent EA-based most cited software for EDM in Computer Science publications from 2007 to 2017

Year	References	Cites	DM problem	EA type	Name	Software type	Language	Licence
2008	Igel, Heidrich-Meisner, and Glasmachers (2008)[a]	85	Clas. & Reg.	GA, ES, MOEA	Shark	Library	C++	GPLv3
2009	Alcala-Fdez et al. (2009)[a,b]	458	Clas., Reg.,	EAs (GA, GP,	KEEL	Suit	Java	GPLv3
			Unsup., Semisup.	DE, MOEA, etc)				
2011	Alcala-Fdez, Alcala, et al. (2011)[a,b]	390	Clas. & Reg.	EAs (GA, GP,	KEEL-Dataset	Suit / Web	Java	GPLv3
				DE, MOEA, etc)				
2016	Vrugt (2016)[a,b]	20	Reg.	DE (BI + DREAM)	DREAM	Toolbox	Matlab	GPLv3

[a] Available software.

[b] Highly cited paper in computer science.

BI = Bayesian inference; Clas. = classification; DE = differential evolution; DREAM = DiffeRential Evolution Adaptive Metropolis; ES = evolutionary strategy; Reg. = regression; Semisup. = semi-supervised; Unsup. = unsupervised.

TABLE 5 Existent EA-based algorithm reviews on EDM in Computer Science publications from 2007 to 2017

Year	References	Cites	Apps	DM problem	Model type	EA type	Algorithm
2010	Fernandez, Garcia, Luengo, Bernado-Mansilla, and Herrera (2010)[a]	59	–	Clas. / Stan. & Imb.	RBS	GBMLs	XCS, CORE, GASSIST, etc
2011	Herrera, Jose Carmona, Gonzalez, and Jose del Jesus (2011)[a]	58	–	SD	RBS & FRBS	MOEA & GA	SDIGA,MESDIF,NMEEF-SD,etc
2012	Barros, Basgalupp, de Acplf, and Freitas (2012)	52	–	Clas. & Reg.	DT	EAs (GA,GP, DE,	LEGAL-Tree, GEA-ODT, etc
						ENN, MOEA, etc)	
2013	Fazzolari, Alcala, Nojima, Ishibuchi, & Herrera (2013)[b]	99	–	Clas., Reg., & Unsup. (ARs)	(Mamdani,TSK&DNF)	MOEA	SPEA2, NSGA-II, PAES,
					FRBS & FARs		MPAES, etc
2014	Mukhopadhyay, Maulik, Bandyopadhyay, and Coello Coello (2014a)	54	–	Clas.	FRBCS, SVM & NN	MOEA	NSGA-II, SPEA2, EMOGA,
							M-PAES, CEMOGA, etc
	Mukhopadhyay, Maulik, Bandyopadhyay, and Coello Coello (2014b	41	–	Reg., SD,	Cluster & AR	MOEA	NSGA-II, SPEA2, PESA-II,
				Unsup. (Clus., ARs), etc			MOGA, MODE, etc
2015	Fernandez, Lopez, Jose del Jesus, and Herrera (2015)[a]	23	–	Clas. (Imb., ML, MI,	FRBS, FCluster	MOEA, GA,	EUSBoost, G3P-MI,
				Ord. & Mon., LQD), Reg., SD,	& FAR	GP, etc	FARLAT-LQD, etc
				DS, Unsup. (Clus., ARs), etc			
2016	Alcala-Fdez and Alonso (2016)[a]	10	–	Clas., Reg., SD, & Unsup.	FRBS, NFS, FAR, etc	EAs & MOEA	FARCHD, FURIA, SLAVE,
							FuzzyApriori, etc

[a] Available software.

[b] Highly cited paper in engineering.

AR = association rules; Clas. = classification; Clus. = clustering; DE = differential evolution; DS = data stream; DT = decision trees; FAR = Fuzzy association rules; FCluster = fuzzy clusters; FRBS = fuzzy rule-based system; Imb. = imbalanced; LQD = low quality data; ML = multi-label; ML = multi-instance; Mon. = monotonic; NFS = neuro-fuzzy system; Ord. = ordinal; RBS = crsip rule-based system; Reg. = regression; SD = subgroup discovery; Stan. = standard; Unsup. = unsupervised.

The first approach we can find in this period (Martens et al., 2007), 2007, makes use of the Ant Colony Optimization (ACO) extending the AntMiner algorithm for multi-class classification by using a better performing MAX-MIN Ant System. The so named AntMiner + is then applied to learn a set of interval rules. In 2008 (Melgani & Bazi, 2008), the authors improved the generalization ability of a SVM by searching for the best value of the parameters that tune its discriminant function based on particle swarm optimization (PSO) in the automatic classification of electrocardiogram beats. The fruit Fly optimization algorithm (FOA) was applied in 2013 (Ze Li, Guo, Jie Li, & Qi Sun, 2013) to annual power load forecasting, where the FOA was used to automatically select the appropriate spread parameter value of a generalized regression NN (GRNN). A biogeography-based optimization (BBO) algorithm was also introduced in 2016 (Yang et al., 2016) to optimize the weights of an SVM for automated classification of brain images.

As said, they are only a representation of the most cited proposals. But the application to DM of many other emergent-EC algorithms can be found among the most cited papers as, the bee colony optimization (BCO), the imperialist competitive algorithm (ICA), the firefly algorithm (FA), etc.

5 | EDM BIBLIOGRAPHICAL ANALYSIS

In this section, we provide a snapshot of the status of publications on EDM according to the ISI WoK, focusing on publications that belong to the top 10% of publications in Computer Science from 2007 to 2017 with the aim of finding out the current research trends in the field. To accomplish this, we first make an analysis of the EDM visibility concerning the number of publications and citations per year. Second, we analyze how the evolutionary techniques used have evolved in the last

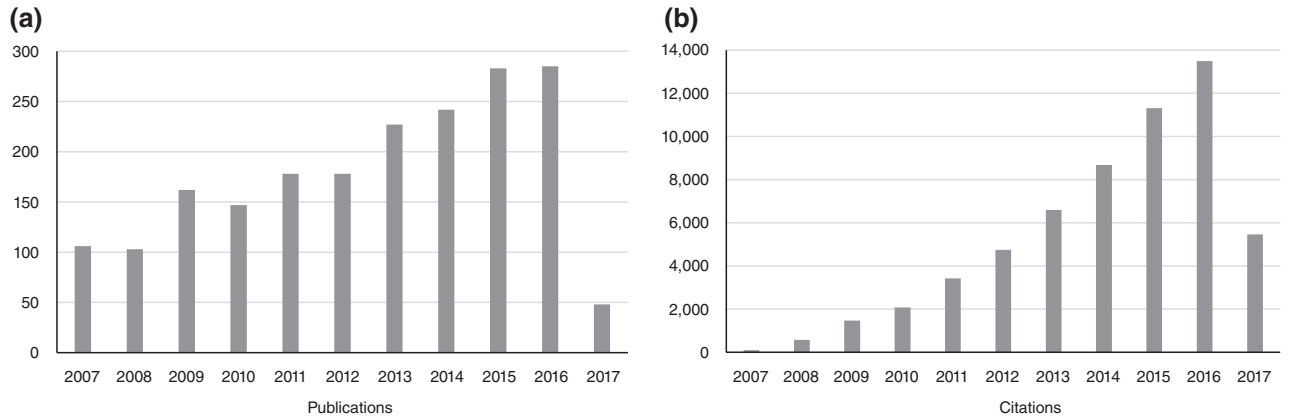

FIGURE 2 Report produced in June 16, 2017; total number of publications on evolutionary data mining (EDM): 1,959; sum of times cited: 57,905; average citations per item: 29.56

10 years, distinguishing between proposals based on EAs and proposals based on other emerging-EC techniques. Finally, we analyze how the publications are distributed among the main ISI categories.

Figure 2 shows the number of published items and the citations per year in the EDM research area that belongs to the top 10% of publications in Computer Science. In Figure 2a, we can see that the number of publications on EDM that appear within the 10% tends to increase each year in general. The exceptions are that the number decreased a little from 2009 to 2010, and remained approximately the same from 2011 to 2012 and from 2015 to 2016. In addition, the number of citations increases substantially over the years (except for the last year since only a few months are considered for this year). All these data allow us to say that the EDM research area is today a mature field with a research community working actively on it.

Table 6 and Figure 3a show the proportion between DM proposals based on EAs and DM proposals based on emerging-EC algorithms that are within the top 10% of the papers in Computer Science for each year. In order to compute the number of EDM papers that are included within the top 10% of the whole Computer Science papers for a year, we have considered the Computer Science table of percentiles provided at the ESI WoS. We take all the papers published in the given year that are over the minimum number of citations for the 10% percentile from both, Q2 and Q1, separately. Let the numbers of taken papers be named as $\#Q2_{10\%}$ and $\#Q1_{10\%}$, respectively. The percentage of EA-based proposals that are in the top 10% with respect to the emerging-EC-based ones in the corresponding year is computed as $\frac{Q1_{10\%}}{Q2_{10\%}} * 100$. And the associated emerging-EC-based percentage with respect to the EA-based ones is computed as $\frac{Q2_{10\%} - Q1_{10\%}}{Q2_{10\%}} * 100$. We can see how the use of EAs with

TABLE 6 Percentages of EDM papers within of the top 10% of papers in Computer Science separated by EA (Q1) and emerging-EC based (EC–EA, ie, Q2−Q1)

		2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
EC	EA-based (%)	85.98	75.00	73.62	66.22	67.60	59.22	56.14	55.14	52.11	51.75	42.86
	Emerging-EC based (%)	14.02	25.00	26.38	33.78	32.40	40.78	43.86	44.86	47.89	48.25	57.14

EA = evolutionary algorithm; EC = evolutionary computation; EDM = evolutionary data mining.

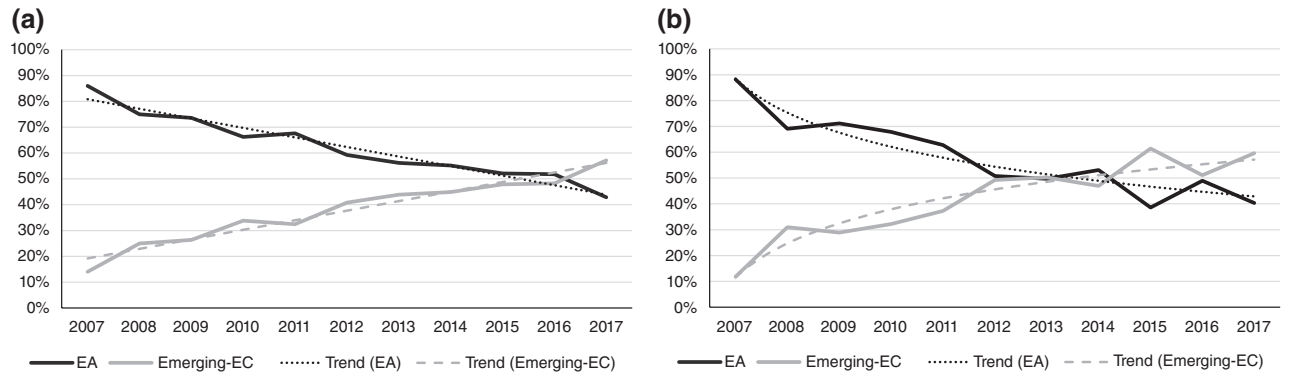

FIGURE 3 (a) Percentages of the number of publications separated by evolutionary algorithm (EA) (Q1) and emerging-evolutionary computation (EC)-based (EC-EA, that is, Q2−Q1); (b) Percentages of citations separated by EA (Q1) and emerging-EC-based (EC-EA, that is, Q2−Q1) percentage of the number of publications and citations in the evolutionary data mining (EDM) area within the top 10% of papers in Computer Science from 2007 to 2017

TABLE 7 Percentage of citations of EDM papers within of the top 10% of papers in Computer Science separated by EA (Q1) and emerging-EC based (EC–EA, ie, Q2−Q1)

		2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
EC	EA-based (%)	88.16	69.07	71.11	67.81	62.75	50.75	49.73	53.06	38.59	48.94	40.35
	Emerging-EC based (%)	11.84	30.93	28.89	32.19	37.25	49.25	50.27	46.94	61.41	51.06	59.65

EA = evolutionary algorithm; EC = evolutionary computation; EDM = evolutionary data mining.

respect to the emerging-EC algorithms decreases linearly over the years within the top 10%, presenting currently almost the same percentage of publications both types.

Table 7 and Figure 3b show the same percentages but in terms of the number of citations received by the so taken proposals. If we consider the number of citations, we can appreciate the same trend, where the percentage of citations of the

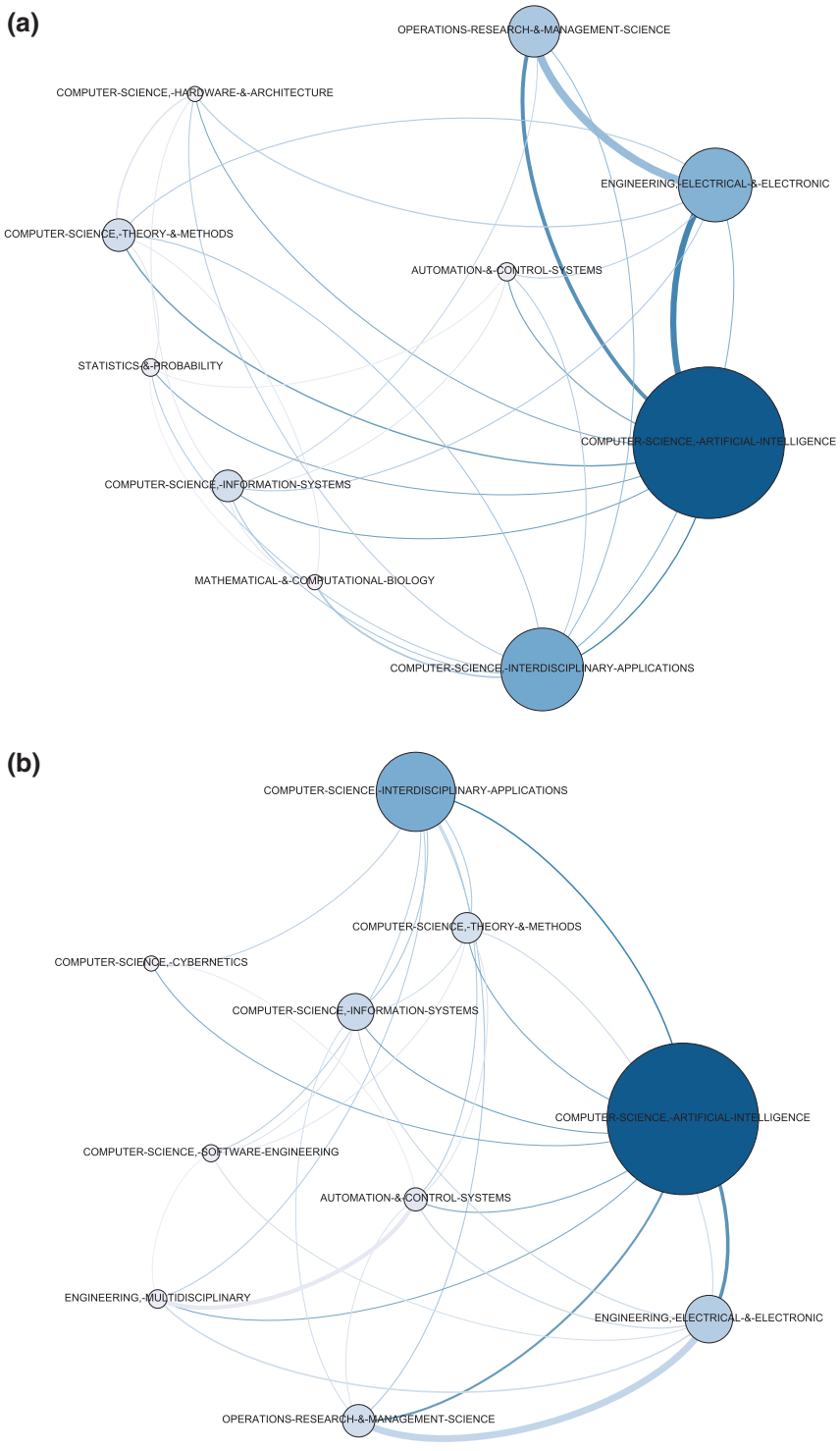

FIGURE 4 Relationships among the most relevant ISI categories for evolutionary data mining (EDM) publications

EA-based papers is even a little lower than the percentage of citations of the EC-based papers from the year 2015. According to these data, we can appreciate how the publication trend is becoming focused on the use of the emergent-EC algorithms.

Finally, to analyze the distribution of the publications among the ISI categories from 2007 to 2017, we have performed a study based on co-concurrence of the publications in the main ISI categories, analyzing the evolution between the periods 2007–2011 and 2012–2017. Notice that the notion of co-concurrence represents the frequency with which publications are simultaneously indexed in two categories. To accomplish this, we have created a visual science map (Moya-Anegón et al., 2004) for each period (see Figure 4), in which each node represents an ISI category (the number of publications is proportional to the darkness and size of the node) and an edge among two nodes represents the degree of co-concurrence of the publications between the two linked categories (the number of co-concurrence is proportional to the darkness and thickness of the edge).[3] The citation counts used to produce Figure 4 were based on all papers returned by query Q2.

These maps (see Figure 4a,b) show a clear view of how publication trends (regarding ISI categories) have evolved in the last 10 years. From the first period (2007–2011), we can appreciate how the categories *Computer Science-Artificial Intelligence*, *Computer Science-Interdisciplinary Applications*, *Engineering Electrical Electronic,* and *Operations Research Management Science* have emerged as the four main categories, with *Computer Science-Interdisciplinary Applications* presenting a higher number of publications than the other categories and sharing a strong link with *Engineering Electrical Electronic* and *Operations Research Management Science.* Moreover, the main categories are linked with other eight categories, of which *Computer Science Theory Methods* and *Computer Science Information Systems* are the most productive.

Although, from the second period (2012–2017), we can appreciate how the two main categories are only *Computer Science-Artificial Intelligence* and *Computer Science-Interdisciplinary Applications* with almost the same number of publications as the previous period, while the categories *Engineering Electrical Electronic* and *Operations Research Management Science* have far fewer number of publications. Moreover, the categories *Statistics Probability*, *Computer Science Hardware Architecture,* and *Mathematical Computational Biology* disappear of the 10 main categories, while the categories *Engineering Multidisciplinary*, *Computer Science Software Engineering,* and *Computer Science Cybernetics* are included in this period.

We can see how EDM is an active area that extends toward new fields. Besides the usual ISI categories (*Computer Science-Artificial Intelligence* and *Computer Science-Interdisciplinary Applications*), several new categories (such as *Engineering-Multidisciplinary*, *Computer Science-Software Engineering,* and *Computer Science-Cybernetics*) represent promising emergent categories that have grown rapidly in the last years, showing a growing interest in application fields and software development.

6 | CRITICAL DISCUSSION AND POSSIBLE FURTHER DIRECTIONS

This section discusses some critical aspects based on the analysis performed in the previous sections as well as some possible further research directions on the studied topic. Several aspects have been found that, from our point of view, could be taken into account when a DM problem is going to be solved in order to properly apply EDM. These issues or recommendations are listed in the following:

- Recommendation—As we have observed in the analysis of the five most cited papers per year, most papers use EAs to enhance well-established ML techniques when they exist for a given type of target problem. EAs only replace well-established ML techniques when this is motivated by the nature of the problem. In our opinion, this is a natural and practical way to apply EDM, since it allows us to get the benefits from the results of extensive research on ML techniques. In fact, as optimization techniques, EAs can easily complement the corresponding DM techniques in order to try to find the optimal learning algorithm input parameters or the initial main structure of the models that are going to be finally obtained by the associated ML algorithm (e.g., NN topology learning before gradient descent parameters learning). In addition, they can be used for tuning of the previously obtained models as a post-processing stage, thus enhancing their performance. Furthermore, they are suitable to replace the typical ML algorithms directly when multiple objectives, complex model structures or particular restrictions need to be considered, since they are highly flexible population-based and fitness-based algorithms, and therefore, they fit very well with these issues.
- Critical issue—As indicated in Garcia et al. (2009), the application of appropriate STs for comparisons is a must. However, we still found many proposals among the most cited ones that do not perform this kind of validation. Even though for particular applications the use of these tests could make no sense, at least the application of the cross-validation method is the expected. However, we found still many proposals applying only the simple HO method. In our opinion, this is a lack of scientific rigor for the EDM proposals that must be solved.

- Critical issue—We have also found the application or adaptation of simple GAs in many of the studied papers. Since there are a lot of proposals on more advanced EAs, this is a real pity to waste their improved search ability. Probably this is the main motivation of the recent success of the emergent-EC algorithms application, since these algorithms were proposed to beat the previous versions of their EA counterparts.

From our point of view, there are still many things to do. In the following, we introduce some potential further directions:

- The application of the emergent-EC algorithms seems a very prolific area. As we can see from the performed study, they are just now outperforming the EA-based approaches in terms of number of highly cited publications and number of citations. This leads to think that further new developments on the EC framework will be successfully applied to the DM problems.
- Further developments for big data mining are still expected. From the analysis performed on the five most cited papers per year, we can see that the big data challenge is just appearing within the most cited contributions on EDM of the last 2 years. Considering the increasingly fast data acquisition from devices, internet activities, etc., it can still be considered as an open problem where new developments on EDM will be still applied.
- The emergent deep learning could be another possible application framework. NNs have been successfully optimized or learned by means of EAs in the specialized literature. As in the case of the big data application, since EAs are very suitable to perform parallelization, they could contribute to enhance or to train the complex and large multi-layer networks required for the deep learning application.
- Application to new types of DM problems. As EAs have successfully solved imbalanced or multi-class classification problems in the last 10 years, the application of EAs could become the easiest way to solve new kinds of complex DM problems that have become more common in the last few years. Multi-variate TS, temporal pattern mining, data privacy, etc., are examples of these relatively new highly complex problems representing a challenge for the traditional ML techniques. Due to their flexibility, EAs or EC algorithms could become a serious alternative to be taken into account.
- Even though it is not a new research direction, we think that problems with multiple objectives are a field where EDM will be still successfully applied. EAs, or in general EC algorithms, are highly suitable and natural tools to evolve a set of solutions representing the different desired trade-offs among the different objectives in a flexible manner. In this way, EDM approaches facing still non-solved real-world applications with a multi-objective nature, or making use of additional objectives in order to better optimize a main objective could still play an important role in the future.

7 | CONCLUSIONS

EDM is a prolific field that takes advantage from the application of EAs in order to enhance or to replace the typical learning algorithms associated with the different problems or DM techniques. EAs are powerful population-based optimization techniques that are able to improve or even learn any type of predictive or descriptive model, evolving their structure and/or definition parameters together. In this review article the most cited contributions on EDM from the last 10 years (including the current year) have been analyzed from two different points of views. First, analyzing the five most cited papers per year. Second, contrasting the classical EA-based approaches with the emergent-EC based ones by focusing on the EDM approaches that can be found within the 10% of the most cited papers from the ISI WoS per year. These top cited papers have been mainly found in the classification and regression areas, addressing a high number of particular applications, imbalanced classification problems, multiple objectives and interpretability issues, and TS forecasting. A number of contributions have been also found on unsupervised learning, addressing mainly clustering problems, and eventually performing AR mining and subgroup discovery.

From this analysis, we can conclude that most of the top cited papers apply EAs in order to enhance the existent ML techniques, only replacing them when the problem nature motivates their use. For example, when multiple objectives should be taken into account searching for a set of solutions with different objective trade-offs or when learning fuzzy systems, since there is no well-established ML technique to efficiently learn these types of rule structures together with their optimal definition parameters. In view of the studied contributions, we recommend the use of more advanced EAs from the specialized literature and the use of STs when it is possible, or at least cross-validation. Moreover, the application of the emergent-EC algorithms has been found to be a prolific current trend and a promising research direction since we have observed that in the last few years they overcome the EAs in the number of publications and citations within the 10% of the most cited papers. Developments for big data, the emergent deep learning and problems with multiple objectives are also proposed as possibly interesting research areas to be explored in a near future.

ACKNOWLEDGMENTS

This work was supported in part by the Spanish Ministry of Economy and Competitiveness under Grants TIN2014-57251-P and TIN2015-68454-R, and the Andalusian Government under Grant P11-TIC-7765.

Conflict of interest

The authors have declared no conflicts of interest for this article.

NOTES

[1]Formerly the Institute for Scientific Information (ISI) Web of Science (WoS): https://apps.webofknowledge.com/.

[2]The Essential Science Indicators (ESI) of the Web of Knowledge (WoK): https://apps.webofknowledge.com/.

[3]The free software Science Mapping Analysis Tool (SciMAT) (Cobo et al., 2012) and The Open Graph Viz Platform (Gephi) have been used to create the maps.

REFERENCES

Ahmadizar, F., Soltanian, K., AkhlaghianTab, F., & Tsoulos, I. (2015). Artificial neural network development by means of a novel combination of grammatical evolution and genetic algorithm. *Engineering Applications of Artificial Intelligence*, *39*, 1–13.

Alatas, B., Akin, E., & Karci, A. (2008). Modenar: Multi-objective differential evolution algorithm for mining numeric association rules. *Applied Soft Computing*, *8*(1), 646–656.

Alavi, A. H., & Gandomi, A. H. (2011). A robust data mining approach for formulation of geotechnical engineering systems. *Engineering Computations*, *28*(3–4), 242–274.

Alcala, R., Ducange, P., Herrera, F., Lazzerini, B., & Marcelloni, F. (2009). A multiobjective evolutionary approach to concurrently learn rule and data bases of linguistic fuzzy-rule-based systems. *IEEE Transactions on Fuzzy Systems*, *17*(5), 1106–1122.

Alcala, R., Jose Gacto, M., & Herrera, F. (2011). A fast and scalable multiobjective genetic fuzzy system for linguistic fuzzy modeling in high-dimensional regression problems. *IEEE Transactions on Fuzzy Systems*, *19*(4), 666–681.

Alcala-Fdez, J., Alcala, R., & Herrera, F. (2011). A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. *IEEE Transactions on Fuzzy Systems*, *19*(5), 857–872.

Alcala-Fdez, J., & Alonso, J. M. A. (2016). Survey of fuzzy systems software: Taxonomy, current research trends, and prospects. *IEEE Transactions on Fuzzy Systems*, *24*(1), 40–56.

Alcala-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., Garcia, S., Sanchez, L., et al. (2011). KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, *17*, 255–287.

Alcala-Fdez, J., Sanchez, L., Garcia, S., del Jesus, M. J., Ventura, S., Garrell, J. M., et al. (2009). KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, *13*(3), 307–318.

Antonio Sanz, J., Bernardo, D., Herrera, F., Bustince, H., & Hagras, H. (2015). A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data. *IEEE Transactions on Fuzzy Systems*, *23*(4), 973–990.

Antonio Sanz, J., Fernandez, A., Bustince, H., & Herrera, F. (2013). IVTURS: A linguistic fuzzy rule-based classification system based on a new interval-valued fuzzy reasoning method with tuning and rule selection. *IEEE Transactions on Fuzzy Systems*, *21*(3), 399–411.

Antonio Sanz, J., Galar, M., Jurio, A., Brugos, A., Pagola, M., & Bustince, H. (2014). Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. *Applied Soft Computing*, *20*, 103–111.

Barros, R. C., Basgalupp, M. P., de ACPLF, C., & Freitas, A. A. (2012). A survey of evolutionary algorithms for decision-tree induction. *IEEE Transactions on Systems, Man and Cybernetics Part C—Applications and Reviews*, *42*(3), 291–312.

Baykasoglu, A., Guellue, H., Canakci, H., & Oebakir, L. (2008). Prediction of compressive and tensile strength of limestone via genetic programming. *Expert Systems with Applications*, *35*(1–2), 111–123.

Bhowan, U., Johnston, M., Zhang, M., & Yao, X. (2013). Evolving diverse ensembles using genetic programming for classification with unbalanced data. *IEEE Transactions on Evolutionary Computation*, *17*(3), 368–386.

Brabazon, A., O'Neill, M., & McGarraghy, S. (2015). *Introduction to evolutionary computing* (pp. 17–20). Berlin, Germany: Springer-Verlag.

Cao, J., Lin, Z., & Huang, G. B. (2012). Self-adaptive evolutionary extreme learning machine. *Neural Processing Letters*, *36*(3), 285–305.

Carmona, C. J., Ruiz-Rodado, V., del Jesus, M. J., Weber, A., Grootveld, M., Gonzalez, P., et al. (2015). A fuzzy genetic programming-based algorithm for subgroup discovery and the application to one problem of pathogenesis of acute sore throat conditions in humans. *Information Sciences*, *298*, 180–197.

Chandra, R., & Zhang, M. (2012). Cooperative coevolution of Elman recurrent neural networks for chaotic time series prediction. *Neurocomputing*, *86*, 116–123.

Chandwani, V., Agrawal, V., & Nagar, R. (2015). Modeling slump of ready mix concrete using genetic algorithms assisted training of artificial neural networks. *Expert Systems with Applications*, *42*(2), 885–893.

Cheng, C. H., Chen, T. L., & Wei, L. Y. (2010). A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting. *Information Sciences*, *180*(9), 1610–1629.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, *63*(8), 1609–1630.

Cpalka, K., Lapa, K., Przybyl, A., & Zalasinski, M. (2014). A new method for designing neuro-fuzzy systems for nonlinear modelling with interpretability aspects. *Neurocomputing*, *135*(SI), 203–217.

Cpalka, K., Rebrova, O., Nowicki, R., & Rutkowski, L. (2013). On design of flexible neuro-fuzzy systems for nonlinear modelling. *International Journal of General Systems*, *42*(6, SI), 706–720.

Das, S., & Sil, S. (2010). Kernel-induced fuzzy clustering of image pixels with an improved differential evolution algorithm. *Information Sciences*, *180*(8), 1237–1256.

Demertzis, K., Iliadis, L., Avramidis, S., & El-Kassaby, Y. A. (2017). Machine learning use in predicting interior spruce wood density utilizing progeny test information. *Neural Computing & Applications*, *28*(3), 505–519.

Derrac, J., Chiclana, F., Garcia, S., & Herrera, F. (2016). Evolutionary fuzzy *k*-nearest neighbors algorithm using interval-valued fuzzy sets. *Information Sciences*, *329*(SI), 144–163.

Ding, Y. S., & Zhang, T. L. (2008). Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: An approach with immune genetic algorithm-based ensemble classifier. *Pattern Recognition Letters*, *29*(13), 1887–1892.

Duchanoy, C. A., Moreno-Armendariz, M. A., Urbina, L., Cruz-Villar, C. A., Calvo, H., & JDJ, R. (2017). A novel recurrent neural network soft sensor via a differential evolution training algorithm for the tire contact patch. *Neurocomputing*, *235*, 71–82.

Eiben, A., & Smith, J. (2003). *Introduction to evolutionary algorithms*. Berlin, Heidelberg: Springer-Verlag.

Fazzolari, M., Alcala, R., Nojima, Y., Ishibuchi, H., & Herrera, F. (2013). A review of the application of multiobjective evolutionary fuzzy systems: Current status and further directions. *IEEE Transactions on Fuzzy Systems*, *21*(1), 45–65.

Fei, S.-W., & Zhang, X. B. (2009). Fault diagnosis of power transformer based on support vector machine with genetic algorithm. *Expert Systems with Applications*, *36*(8), 11352–11357.

Fernandez, A., Garcia, S., Luengo, J., Bernado-Mansilla, E., & Herrera, F. (2010). Genetics-based machine learning for rule induction: State of the art, taxonomy, and comparative study. *IEEE Transactions on Evolutionary Computation*, *14*(6), 913–941.

Fernandez, A., Lopez, V., Jose del Jesus, M., & Herrera, F. (2015). Revisiting evolutionary fuzzy systems: Taxonomy, applications, new trends and challenges. *Knowledge-Based Systems*, *80*(SI), 109–121.

Gandomi, A. H., & Alavi, A. H. (2011). Multi-stage genetic programming: A new strategy to nonlinear system modeling. *Information Sciences*, *181*(23), 5227–5239.

Gandomi, A. H., & Alavi, A. H. (2012a). A new multi-gene genetic programming approach to nonlinear system modeling. Part I: Materials and structural engineering problems. *Neural Computing & Applications*, *21*(1), 171–187.

Gandomi, A. H., & Alavi, A. H. (2012b). A new multi-gene genetic programming approach to non-linear system modeling. Part II: Geotechnical and earthquake engineering problems. *Neural Computing & Applications*, *21*(1), 189–201.

Garcia, S., Fernandez, A., Luengo, J., & Herrera, F. (2009). A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability. *Soft Computing*, *13*(10), 959–977.

Ghosh, A., Mishra, N. S., & Ghosh, S. (2011). Fuzzy clustering algorithms for unsupervised change detection in remote sensing images. *Information Sciences*, *181*(4), 699–715.

Gorzalczany, M. B., & Rudzinski, F. (2016). A multi-objective genetic optimization for fast, fuzzy rule-based credit classification with balanced accuracy and interpretability. *Applied Soft Computing*, *40*, 206–220.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). San Francisco, CA: Morgan Kaufmann.

Handl, J., & Knowles, J. (2007). An evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation*, *11*(1), 56–76.

Herrera, F., Jose Carmona, C., Gonzalez, P., & Jose del Jesus, M. (2011). An overview on subgroup discovery: Foundations and applications. *Knowledge and Information Systems*, *29*(3), 495–525.

Huang, C. F. (2012). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, *12*(2), 807–818.

Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856.

Igel, C., Heidrich-Meisner, V., & Glasmachers, T. (2008). Shark. *Journal of Machine Learning Research*, *9*, 993–996.

Ishibuchi, H., & Nojima, Y. (2007). Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning. *International Journal of Approximate Reasoning*, *44*(1), 4–31.

Jose Gacto, M., Alcala, R., & Herrera, F. (2010). Integration of an index to preserve the semantic interpretability in the multiobjective evolutionary rule selection and tuning of linguistic fuzzy systems. *IEEE Transactions on Fuzzy Systems*, *18*(3), 515–531.

Kisi, O., Shiri, J., & Tombul, M. (2013). Modeling rainfall-runoff process using soft computing techniques. *Computers & Geosciences*, *51*, 108–117.

Krawczyk, B., Galar, M., Jelen, L., & Herrera, F. (2016). Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing*, *38*, 714–726.

Krawczyk, B., Wozniak, M., & Schaefer, G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, *14*(C), 554–562.

Kuang, F., Xu, W., & Zhang, S. (2014). A novel hybrid KPCA and SVM with GA model for intrusion detection. *Applied Soft Computing*, *18*, 178–184.

Lei, Y., Zuo, M. J., He, Z., & Zi, Y. (2010). A multidimensional hybrid intelligent method for gear fault diagnosis. *Expert Systems with Applications*, *37*(2), 1419–1430.

Mansoori, E. G., Zolghadri, M. J., & Katebi, S. D. (2008). SGERD: A steady-state geneticalgorithm for extracting fuzzy classification rules from data. *IEEE Transactions on Fuzzy Systems*, *16*(4), 1061–1071.

Martens, D., Backer, M. D., Haesen, R., Vanthienen, J., Snoeck, M., & Baesens, B. (2007). Classification with ant colony optimization. *IEEE Transactions on Evolutionary Computation*, *11*(5), 651–665.

Martinez, R., Castillo, O., & Aguilar, L. T. (2009). Optimization of interval type-2 fuzzy logic controllers for a perturbed autonomous wheeled mobile robot using genetic algorithms. *Information Sciences*, *179*(13), 2158–2174.

Melgani, F., & Bazi, Y. (2008). Classification of electrocardiogram signals with support vector machines and particle swarm optimization. *IEEE Transactions on Information Technology in Biomedicine*, *12*(5), 667–677.

Menendez, H. D., Barrero, D. F., & Camacho, D. A. (2014). Genetic graph-based approach for partitional clustering. *International Journal of Neural Systems*, *24*(3), 1430008.

Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., & Muñoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, *61*(1), 129–145.

Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., & Coello Coello, C. A. (2014a). A survey of multiobjective evolutionary algorithms for data mining: Part I. *IEEE Transactions on Evolutionary Computation*, *18*(1, SI), 4–19.

Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., & Coello Coello, C. A. (2014b). Survey of multiobjective evolutionary algorithms for data mining: Part II. *IEEE Transactions on Evolutionary Computation*, *18*(1, SI), 20–35.

Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015). Classification of healthcare data using genetic fuzzy logic system and wavelets. *Expert Systems with Applications*, *42*(4), 2184–2197.

Oliveira, G. V., Coutinho, F. P., Campello, R. J. G. B., Naldi, M. C., & Soc Brasileira Comp SBC, Univ Federal do Rio Grande do Norte (2017). Improving *k*-means through distributed scalable metaheuristics. *Neurocomputing*, *246*(SI), 45–57 4th Brazilian Conference on Intelligent Systems (BRACIS), Natal, Brazil, Nov 4–7, 2015.

Pulkkinen, P., & Koivisto, H. (2010). A dynamically constrained multiobjective genetic fuzzy system for regression problems. *IEEE Transactions on Fuzzy Systems*, *18*(1), 161–177.

Rojas, I., Valenzuela, O., Rojas, F., Guillen, A., Herrera, L. J., Pomares, H., … Pasadas, M. (2008). Soft-computing techniques and ARMA model for time series prediction. *Neurocomputing*, *71*(4–6), 519–537.

Rudzinski, F. (2016). A multi-objective genetic optimization of interpretability-oriented fuzzy rule-based classifiers. *Applied Soft Computing*, *38*, 118–133.

Serdio, F., Lughofer, E., Zavoianu, A. C., Pichler, K., Pichler, M., Buchegger, T., & Efendic, H. (2017). Improved fault detection employing hybrid memetic fuzzy modeling and adaptive filters. *Applied Soft Computing*, *51*, 60–82.

Shao, Y. H., Wang, Z., Chen, W. J., & Deng, N. Y. (2013). A regularization for the projection twin support vector machine. *Knowledge-Based Systems*, *37*, 203–210.

Shen, C. P., Lin, J. W., Lin, F. S., Lam, A. Y. Y., Chen, W., Zhou, W., … Lai, F. (2017). GA-SVM modeling of multiclass seizure detector in epilepsy analysis system using cloud computing. *Soft Computing*, *21*(8), 2139–2149.

Suresh, S., Babu, R. V., & Kim, H. J. (2009). No-reference image quality assessment using modified extreme learning machine classifier. *Applied Soft Computing*, *9*(2), 541–552.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston, MA, USA: Pearson Addison Wesley.

Triguero, I., Gonzalez, S., Moyano, J. M., Garcia, S., Alcala-Fdez, J., Luengo, J., et al. (2017). KEEL 3.0: An open source software for multi-stage analysis in data mining. *International Journal of Computational Intelligence Systems*, *10*, 1238–1249.

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling & Software*, *75*(SI), 273–316.

Wang, L., Wang, Z., & Liu, S. (2016). An effective multivariate time series classification approach using echo state network and adaptive differential evolution algorithm. *Expert Systems with Applications*, *43*, 237–249.

Wu, C. H., Tzeng, G. H., Goo, Y. J., & Fang, W. C. (2007). A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Systems with Applications.*, *32*(2), 397–408.

Xing, B., & Gao, W. J. (2014). *Innovative computational intelligence: A rough guide to 134 clever algorithms*. Cham, Switzerland: Springer.

Yang, G., Zhang, Y., Yang, J., Ji, G., Dong, Z., Wang, S., … Wang, Q. (2016). Automated classification of brain images using wavelet-energy and biogeography-based optimization. *Multimedia Tools and Applications*, *75*(23), 15601–15617.

Yang, X. S. (2014). Recent advances in swarm intelligence and evolutionary computation. In *Studies in computational intelligence* (Vol. *585*). Cham, Switzerland: Springer.

Yao, J., Vasilakos, A., & Pedrycz, W. (2013). Granular computing: Perspectives and challenges. *IEEE Transactions on Cybernetics*, *43*(6), 1977–1989.

Ze Li, H., Guo, S., Jie Li, C., & Qi Sun, J. (2013). A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm. *Knowledge-Based Systems*, *37*, 378–387.

Zhu, Z., Ong, Y. S., & Dash, M. (2007). Wrapper-filter feature selection algorithm using a memetic framework. *IEEE Transactions on Systems, Man and Cybernetics Part B-Cybernetics*, *37*(1), 70–76.

How to cite this article: Alcalá R, Gacto MJ, Alcalá-Fdez J. Evolutionary data mining and applications: A revision on the most cited papers from the last 10 years (2007–2017). *WIREs Data Mining Knowl Discov*. 2017;e1239. https://doi.org/10.1002/widm.1239