Big Data Analytics

RESEARCH Open Access

Study on the use of different quality measures within a multi-objective evolutionary algorithm approach for emerging pattern mining in big data environments

Ángel Miguel García-Vico[1*†], Pedro González[1†], Cristóbal José Carmona[1,2†] and María José del Jesus[1]

*Correspondence: agvico@ujaen.es
†Equal contributors
†Ángel Miguel García-Vico, Pedro González and Cristóbal José Carmona contributed equally to this work.
[1]Department of Computer Science, University of Jaén, Paraje Las Lagunillas, s/n, 23071 Jaén, Spain
Full list of author information is available at the end of the article

Abstract

Background: Emerging pattern mining is a data mining task that extracts rules describing discriminative relationships amongst variables. These rules should be understandable for the experts. Comprehensibility of a rule is traditionally determined by several objectives, which can be calculated by different measures. In this way, multi-objective evolutionary algorithms are suitable for this task. Currently, the growing amount of data makes traditional data mining tasks unable to process them in a reasonable time. These huge amounts of data make even more interesting the extraction of rules that can easily describe the underlying phenomena of this big data. So far there is only one algorithm for emerging pattern mining developed based on multi-objective evolutionary algorithms for big data, the BD-EFEP algorithm. The influence of the selection of different quality measures as objectives in the search process is analysed in this paper.

Results: The results show that the use of the combination based on Jaccard index and false positive rate is the one with the best trade-off for descriptive induction of emerging patterns.

Conclusions: It is recommended the use of this combination of quality measure as optimisation objectives in future multi-objective evolutionary algorithm developments for emerging pattern mining focused in big data.

Keywords: Evolutionary algorithms, Fuzzy systems, Big data, Emerging pattern mining

Introduction

The amount of information generated everyday has suffered an exponential growth since the last decades. Nowadays, it is estimated that the Internet generates almost 57 TB of traffic in one second: 8000 tweets are posted, almost 900 Instagram pictures are uploaded, more than 2.5 million emails are sent, etc. [1]. These huge amount of data is commonly called big data [2]. Big data is not only related to the amount of data but is also related to the variety of sources where these data come from and the arrival velocity into the system. This amount of data should be analysed for the extraction of valuable knowledge that can ease the decision making processes. However, the extraction of knowledge within these huge amounts of data is not possible by means of traditional data mining techniques. One of the most popular frameworks to deal with big data is MapReduce [3, 4], where its open source implementations Hadoop [5] and Spark [6] are the most popular ones. MapReduce consists in a distributed processing system using a divide-and-conquer approach.

The emerging pattern mining (EPM) [7, 8] is a data mining task within the supervised descriptive rule discovery (SDRD) framework [9] whose main aim is the description of discriminative characteristics, i.e., the values of the variables in data, amongst classes of a dataset or the description of emerging phenomena in data that arrives continuously to the system. The knowledge extracted is usually represented by means of rules. These are usually useful for the experts when they are understandable for them. In fact, the comprehensibility of these rules is even more important in big data environments, because of the relevance of the understanding of the underlying phenomena in such a complex environment. The EPM task has been successfully applied in several fields such as chemistry [10, 11], medicine [12, 13], tourism management [14], photovoltaic technology [15], among others [16].

The comprehensibility of a rule is a subjective measure that depends on the expert and the problem analysed. However, a rule within a SDRD task such as EPM is determined as comprehensible if it covers a huge amount of positive examples, it is reliable and the knowledge described is not obvious for the expert, i.e., it is interesting [9]. As it is a multi-objective problem, one of the most relevant algorithms developed within the EPM task focused on the extracion of knowledge within big data environments is a multi-objective evolutionary algorithim called BD-EFEP [17]. This method, developed for the Spark framework, can extract relevant and quality knowledge within a reasonable amount of time. Nevertheless, the set of quality measures selected as optimisation objectives within the evolutionary process has a significant effect on the quality of the results extracted as they are used to guide the search in the evolutionary process.

In this paper, a study about the behaviour of different combinations of quality measures used as optimisation objectives in the BD-EFEP algorithm for a set of big data problems is presented. For this purpose, the paper is organised as follows. First, a brief description of the main concepts presented in this paper is presented in the "Background" section. Next, the methodology used to achieve this objective and the experimental framework are outlined. After that, the results extracted and an analysis of them are shown. Finally, the conclusions of this work are presented.

Background

A brief description of the main concepts introduced in this paper is presented in this section. Firstly, the EPM task is described. Secondly, a summary of quality measures classified by the main objectives of EPM models is depicted. Finally, the MapReduce paradigm and the main algorithms developed for EPM under this approach are briefly outlined.

Emerging pattern mining

EPM is a data mining task for searching patterns whose support significantly differs between two classes or databases [7, 8]. A pattern is considered as emerging if and only if its growth rate (GR) is larger than a given threshold $\rho > 1$. The GR measure is defined as in Eq. 1.

$$GR(x) = \begin{cases} 0, & IF\ Sup_{D_1}(x) = Sup_{D_2}(x) = 0, \\ \infty, & IF\ Sup_{D_1}(x) \neq 0 \wedge Sup_{D_2}(x) = 0, \\ \frac{Sup_{D_1}(x)}{Sup_{D_2}(x)}, & another\ case \end{cases} \quad (1)$$

where $Sup_{D_i}(x)$ is the support of the pattern x in dataset i (D_i).

The main objectives of EPM are the discovery of significant differences amongst classes, emerging trends throughout time or the detection of differences amongst variables. This work is mainly focused on the former objective. Patterns that are able to correctly describe the underlying phenomena in data are searched.

These patterns can be represented as rules in the form [18]:

$$R : Cond \rightarrow Class \quad (2)$$

where *Cond* is the antecedent part of the rule and *Class* is the consequent part of the rule that represents the variable of interest.

There are several algorithms developed for EPM throughout the literature. These can be classified according to the approach used to mine the emerging patterns (EPs). A complete review of these approaches can be found in [8]. Nevertheless, the majority of the algorithms developed along the literature has been focused on classification purposes. These models are usually very hard to understand because they tend to extract a high number of rules with a high number of variables in order to obtain the maximum classification accuracy.

Actually, the objective of classification tasks is the prediction of the value of a variable of interest on unseen instances. In this way, rules present dependencies between them in order to perform a proper classification. As an example, one of the most popular classification methods for EPM models is CAEP [19]. This method performs a prediction on a new example by means of an aggregation of the supports of all the rules that match (or cover) the example. A rule covers an example if the antecedent of the rule matches the example. After that, it assigns the label of the most supported class. Therefore, it can be observed that all the rules that covers the new instance take part in the prediction process. So the underlying phenomena that define the behaviour of the dataset is not clearly defined because there are several rules participating in the classification process. However, EPM tries to describe the emerging behaviour or the discriminative characteristics, so rules can be analysed as independent pieces of knowledge by the expert. In this way, knowledge extracted in EPM should be simple, in terms of low number of variables, with high coverage of the positive class and low error rate. Nevertheless, it is important to remark that it is not necessary to find rules with zero error; rules with low error rate but simpler are desirable as the expert is finding an easy description.

Quality measures used for the description of the objectives

EPM is a descriptive data mining task for finding discriminative rules that correctly describes the underlying phenomena in data. In addition, these rules should be easily comprehensible and interesting for the expert. These characteristics allow us to define several objectives that the rule models extracted in EPM should have [8, 18]:

- Generality. A rule is more general when the number of covered examples is increased. In this way, more general rules allow the extraction of a more comprehensible knowledge, because of the use of less rules to cover the space. So more general rules are desirable.

- Interest. The generality of these rules does not imply that the knowledge extracted should be obvious for the expert. In fact, the expert uses data mining in order to extract new relevant, useful knowledge. The relevance could be the description of an unusual, emerging or discriminative behaviour. In this way, the extraction of novel rules is key.
- Reliability. The main objective of EPM is the extraction of rules that correctly describes the discriminating characteristics of the different classes of the problem. If the problem is not correctly described then the cost could be high. So the knowledge extracted should be as accurate as possible.
- Comprehensibility. The extraction of knowledge that is simple is key in an EPM model. A simple model allows a better learning and use of the knowledge in the decision making process performed by the expert. Therefore, it is important to remark that it is not necessary the extraction of rules with the maximum accuracy; rules with good one but much simpler are preferred.

These objectives are normally used when experts are not available for the determination of the quality of a rule set. These objectives can be calculated by means of different quality measures developed throughout the literature. The majority of these measures are related to some statistical properties regarding to the coverage of the extracted rule. A rule covers an example if the antecedent of the rule matches the example. In addition, the example is correctly covered by the rule if its consequent part matches the class of the example. This measures can be computed by means of a contingency table as depicted in Table 1.

Table 1 presents the contingency table of a rule where p is the number of correctly covered examples, n is the number of incorrectly covered examples, $\overline{p}$ is the number of incorrectly non-covered examples, and $\overline{n}$ is the number of correctly non-covered examples.

As mentioned previously, there are several objectives that should be accomplished in order to extract high quality EPM rule models. Following the results presented in [20], where a set of quality measures were analysed in order to determine their correlation, the results of the study showed that the analysed quality measures present a low average correlation amongst them. According to this study, and the studies presented in [8, 18] the most interesting quality measures for EPM are depicted below:

- Measures for generality. They try to determine the generality of the rule. The most used measures are:
  - True Positive Rate (TPR). It computes the ratio of correctly covered examples with respect to the total amount of positive examples. It is calculated as [21]:

$$TPR(R) = \frac{p}{p + \overline{p}} \tag{3}$$

Table 1 Contingency table of a rule

	Predicted condition	
True condition	Positive	Negative
Positive	$p = tp$	$\overline{p} = fn$
Negative	$n = fp$	$\overline{n} = tn$

  - Support Difference (SupDiff). It measures the difference between the TPR and the ratio of examples incorrectly covered. It is computed as [22]:

$$SupDiff(R) = \frac{p}{p+\overline{p}} - \frac{n}{n+\overline{n}} \tag{4}$$

- Measures for interest. These measures try to determine the relevance of a rule for the expert. The most used are:
  - Weighted Relative Accuracy (WRAcc). It estimates the trade-off between the generality of the rule and its accuracy gain. It is calculated as [18]:

$$WRAcc(R) = \frac{p+n}{p+n+\overline{n}+\overline{p}} \left(\frac{p}{p+n} - \frac{p+\overline{p}}{p+n+\overline{n}+\overline{p}} \right) \tag{5}$$

  - Jaccard index (Jac). It calculates the similitude between two datasets. In this case they are the set of positive examples and the set of examples covered by the rule. It is measured as [23]:

$$Jac(R) = \frac{p}{p+\overline{p}+n} \tag{6}$$

- Measures for reliability. These measures attempt to determine the accuracy of the rule. The most used are described below.
  - Growth Rate (GR). It is the measure that defines an EP. It measure the ratio of the supports of one class with respect to the remaining. It is computed as [7]:

$$GR(R) = \begin{cases} 0, & IF\ \frac{p}{p+\overline{p}} = \frac{n}{n+\overline{n}} = 0, \\ \infty, & IF\ \frac{p}{p+\overline{p}} \neq 0 \wedge \frac{n}{n+\overline{n}} = 0, \\ \frac{p(n+\overline{n})}{n(p+\overline{p})}, & another\ case \end{cases} \tag{7}$$

  - Confidence (Conf). It defines the ratio of the predictive capacity of the rule for the positive class with respect to the examples it covers. It is computed as [24]:

$$Conf(R) = \frac{p}{p+n} \tag{8}$$

  - False Positive Rate (FPR). It determines the percentage of incorrectly covered examples with respect to the total amount of negative examples. This measure should be minimised in order to extract reliable rules. It is computed as [25]:

$$FPR(R) = \frac{n}{n+\overline{n}} \tag{9}$$

  - Geometric mean TPR-TNR (G-mean). It quantifies the trade-off between the accuracy of a rule with respect to positive and negative examples as the product of the TPR and the true negative rate (TNR). It is calculated as [26]:

$$G-mean(R) = \sqrt{\frac{p}{p+\overline{p}} \cdot \frac{\overline{n}}{n+\overline{n}}} \tag{10}$$

- Measures for comprehensibility. These measures quantify the simplicity of the rule set extracted. The most used are:
  - Number of rules (n_r). It quantifies the number of rules extracted. A simple model has a low number of rules.
  - Number of variables (n_v). It measures the average number of variables that each rule contains.

Although the measures presented have been classified within a single objective, some measures have a hybrid behaviour, i.e., they can catch elements of more than one objective. The measures presented together the objectives they can calculate are shown in Table 2 highlighted with a X.

Big data in emerging pattern mining

One of the main issues related to EPM is the complexity of the extraction of the EPs. It is exponential with respect to the number of variables of the problem [27]. Throughout the literature, researchers focused their efforts in the development of restrictions which allow the extraction of subsets of high quality EPs. In addition, methods has been developed to efficiently extract EPs from these subgroups [8], where the use of evolutionary algorithms is highlighted in the recent years due to the descriptive quality of the extracted patterns [15, 17]. Nevertheless, it is necessary the development of more efficient approaches because of the huge amounts of data generated everyday. Nowadays, the most popular paradigm to deal with huge amounts of data is MapReduce [4, 28, 29]. It is based on the divide-and-conquer programming paradigm and it allows an easy parallel execution throughout several machines. Actually, the mechanisms that allow the parallelism are transparent to the developer.

MapReduce contains two main phases:

- Map phase: The master node of a cluster of computers creates a partition of the data. Each chunk of data is sent to a worker node of the cluster. Then, each node compute a result with respect to the data it owns. This result is sent to the master node when the work is finished.
- Reduce phase: The master node collects all the results from the worker nodes, it combines them, and it returns the final result of the problem to the expert.

One of the most popular frameworks implementing MapReduce is Spark [6]. Spark is a high-efficiency computing framework for the processing of massive amounts of data. Its popularity is due to the intensive use of main memory which is very efficient on iterative algorithms.

There development of evolutionary algorithms for Big Data is a challenge because of its complexity. Nowadays, there are several efforts for the development of evolutionary algorithms in several data mining tasks [30–33]. To the best of our knowledge, the developed algorithms are EvAEFP-Spark [34] and BD-EFEP [17]. Both methods use a global

Table 2 Classification of the most important quality measures for emerging pattern mining

Quality measure		Comprehensibility	Generality	Reliability	Interest
TPR (3)	True positive rate		X		
SupDiff (4)	Support difference		X	X	
WRAcc (5)	Weighted relative accuracy		X	X	X
Jac (6)	Jacard index		X	X	X
GR (7)	Growth rate			X	
Conf (8)	Confidence			X	
FPR (9)	False positive rate			X	
$G - mean$ (10)	Geometric mean TPR-TNR		X	X	
n_r	Number of rules	X			
n_v	Number of variables	X			

MapReduce approach which allows the extraction of the same results regardless the number of partitions used. A general schema of the approach of both methods is presented in Fig. 1. MapReduce is used on the evaluation of individuals. As previously mentioned, the measures used in the evaluation of the individuals can be calculated by means of a contingency table. In the map phase, for each individual, a contingency table is calculated on each worker node with the data it owns. The reduce phase joins these tables in order to get the final one for each individual. Finally, the objective measures are calculated by means of this contingency table.

The main features of both methods are depicted below:

- EvAEFP-Spark is based on a mono-objective evolutionary algorithm. It uses a "chromosome = rule" representation where only the antecedent part of the rule is represented, so only rules for one class are extracted on a single execution. It follows an iterative rule learning (IRL) approach [35] where only the best rule is returned at the end of the evolutionary process. After that, if the stopping criteria is not fulfilled, the evolutionary process is started again in order to find another rule. The algorithm stops when the rule extracted by the evolutionary process is not an EP or it does not cover examples not covered by the previously extracted rules.
- BD-EFEP is a multi-objective evolutionary algorithm (MOEA) based on the NSGA-II approach following a competitive-cooperative approach [36]. In this approach, the individuals compete with each other by means of a token competition operator [37]. They cooperate by means of the genetic operators. In addition, it uses a "chromosome = rule" representation where the whole rule is represented, which leads to the extraction of rules for all classes in a single execution.

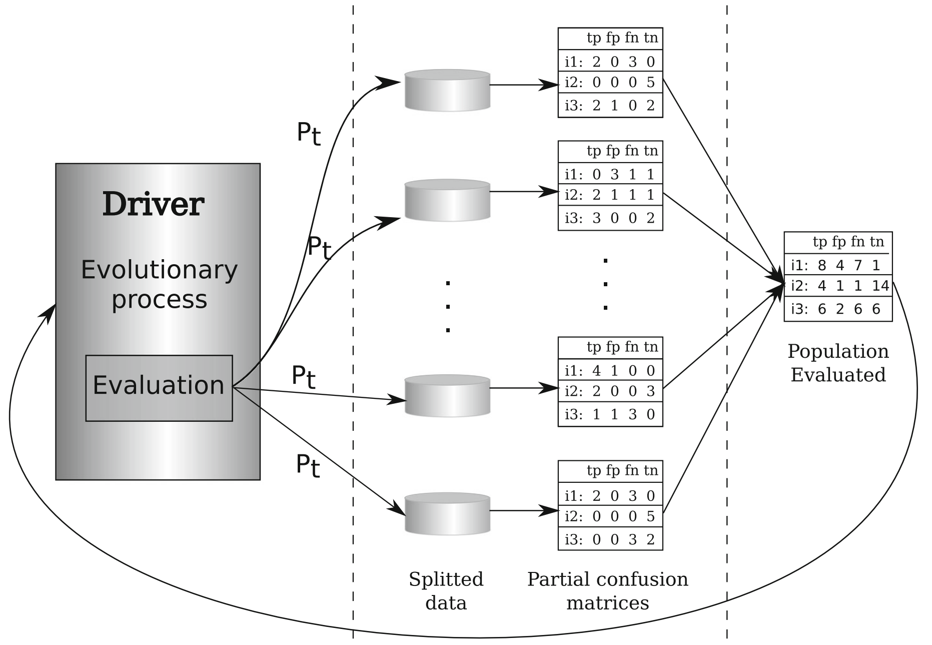

Fig. 1 MapReduce. MapReduce schema of both BD-EFEP and EvAEFP-Spark algorithms. From left to rigth, the driver runs the evolutionary process until the evaluation of the individuals, where a MapReduce job is triggered. Next, the driver sends the population to each worker node within the map phase, where a contingency table is calculated for each individual with respect to the data in that worker. Finally, the reduce phase joins all these tables to obtain the final contingency table of each individual

Methods

This section presents the methods and the experimental framework used to carry out the experimental study. The main aim is the determination of the most suitable combination of quality measures for the extraction of comprehensible EPM rules by means of a MOEA algorithm in big data environments.

Initial hypotheses

As mentioned before, EPM needs the optimisation of several objectives in order to extract a set of rules with high discriminative power and easily comprehensible by the expert. Nevertheless, some of these objectives are conflicting. As an example, the extraction of a more general rule normally decreases its reliability [38]. Due to the conflicting nature of these objectives, a multi-objective optimisation algorithm is well-suited for the extraction of EPs with a good trade-off amongst the objectives.

To the best of our knowledge, there is only one algorithm developed so far based on multi-objective optimisation in EPM for Big Data environments: the BD-EFEP algorithm [17]. It has been demonstrated that the performance of a MOEA algorithm decreases when the number of optimisation objectives is greater than two [39]. Therefore, in order to avoid the loss of performance in BD-EFEP it is necessary the use of hybrid measures that can bind these objectives together. It is important to remark that the measures used must reflect a conflicting nature to maximise the performance of the algorithm.

Following these antecedents, the initial hypothesis of this work focuses on two aspects. On the one hand, we question whether the conflicting nature of the objectives used in EPM is more evident within big data environments, making multi-objective optimisation necessary. On the other hand, we consider determining which of those quality measures better represents the objectives of EPM without degrading the performance of the BD-EFEP algorithm.

Datasets

The study is carried out using a set of 6 well-known large-scale real datasets from the UCI reprository [40]. These datasets follow the ARFF file format of WEKA [41], where the range of instances is up to eleven million instances. The number of variables is low. However, the dimension of datasets employed is enough to be addressed by big data methods. Moreover, the applications of the datasets employed belong to different areas such as the determination of the income according to census data in *census*, network attacks in *Kddcup*, epidemiological cancer in *rlcp* and particle physics in *Susy*, *higgs* and *hepmass*. In this study, datasets were splitted by means of a five-fold cross-validation scheme.

The properties of these datasets are presented in Table 3, where the number of examples (# Instances), the number of variables (# Variables), separated in real, integer and nominal (R/I/N), the size of the datasets in gigabytes (GB), and the number of classes (# Classes) are shown.

Algorithms and parameters

To the best of our knowledge, there are two algorithms for the extraction of EPs within big data environments, the EvAEFP-Spark [34] and the BD-EFEP [17] algorithms. In this study, only the results of BD-EFEP are analysed. EPM is a many objective problem and the use of a mono-objective algorithm such as EvAEFP-Spark can limit the conclusions

Table 3 Properties of the datasets used in the experiments

Name	# Instances	# Variables (R/I/N)	Size (GB)	# Classes
census	299284	41 (1/12/28)	0.151	2
kddcup	494020	41 (26/0/15)	0.049	23
rlcp	5749132	11 (11/0/0)	0.452	2
susy	5000000	18 (18/0/0)	1.503	2
higgs	11000000	28 (28/0/0)	4.772	2
hepmass	10500000	29 (29/0/0)	4.886	2

extracted in this work. In addition, the EvAEFP-Spark algorithm is not able to extract results on the *hepmass* and *higgs* datasets within 24 hours. The parameters used for BD-EFEP are the ones proposed by the authors. In particular, the parameters are: number of labels = 3; number of evaluation = 10000; population length = 51; crossover probability = 0.6 and mutation probability = 0.1.

Rule selection

Traditionally, MOEA algorithms return a set of non-dominated solutions which is called the Pareto front. In EPM, the objective is to find a set of rules able to describe the discriminative characteristics between the classes. Each rule in EPM is an individual piece of knowledge. Therefore, the selection criterion for the final set of rules presented to the expert is the extraction of the whole Pareto front. This allow the expert to give them the choice, according to their objectives and their experience, of the most important ones.

Quality measures analysed

It is necessary to distinguish between the quality measures used as search objectives in the BD-EFEP algorithm and the quality measures employed for the analysis of the rules extracted by them. In this way, we can use all the measures previously presented as a search objective. Nevertheless,according to [39], the use of more than two optimisation objectives decrease the performance of this kind of methods. Therefore, only combinations of two quality measures as optimisation objectives are used. Additionally, the optimisation objectives should be able to cover all EPM objectives and should be conflicting among them in order to achieve the best performance. Following these premisses, 8 combinations have been analysed:

- Jac and TPR.
- G-mean and Jac.
- Jac and FPR.
- G-mean and WRAcc.
- Jac and WRAcc.
- SupDiff and Jac.
- TPR and FPR.
- WRAcc and SupDiff.

It is important to remark that GR was not used as optimisation objective because, as demonstrated in [18], WRAcc is able to obtain patterns with high GR and with the best trade-off between its generality and its reliability.

On the other hand, the analysis of the performance of each combination of measures can be determined by means of any measure. Nevertheless, researchers usually employ a subset of measures. In this study, the measures employed are the ones presented in [8], i.e., n_r, n_v, WRACC, CONF, GR, TPR and FPR as they are one of the most relevant for the determination of the descriptive quality of emerging patterns. It is important to remark that, although GR is the measure that define an emerging pattern, this measure is not able to show us other important aspects in EPM such as the generality of the patterns and the amount of error rate produced. Therefore, it is necessary the use of other measures such as TPR, FPR for the determination of those aspects. Moreover, n_r and n_v allow us the determination of the complexity of the models extracted.

Results and discussion

The comparison of the different combinations of quality measures in order to be used as objectives for BD-EFEP is presented in Table 4. The average results are presented on each row. A highlighted value represents the best value for each analysed quality measure, i.e., the best one for each column.

Table 4 presents the average results obtained by different combinations of quality measures. Figure 2 presents the results in Table 4 in a graphical way. It is important the remark that values in this figure were normalised in order to ease the comparison. According to the results presented, an analysis of the quality of the rules extracted based on the measurements proposed in [8] is presented below:

- n_r. The number of rules extracted for BD-EFEP for each combination is, in general, acceptable. Jac and TPR are highlighted as the objectives whose number of rules extracted is the lowest. This behaviour is due to these objectives are mainly oriented towards generality. So less rules are necessary to cover the search space.
- n_v. The average number of variables is also acceptable in all the cases. The best result is for the combination of Jac and TPR and it was justified previously.
- WRACC. For this measure, the best result is obtained by the combination Jac and FPR. The main objective of WRACC measure is to find a trade-off between the generality and the accuracy of the rule. In this way, Jac tries to find rules as general and accurate as possible, while FPR tries to find the most accurate ones. This produces the extraction of rules with a good balance between generality and reliability.

Table 4 Average results obtained by BD-EFEP with different combinations of objective measures for emerging pattern mining

Combination	n_r	n_v	WRACC	CONF	GR	TPR	FPR
Jac & TPR	**11.167**	**2.811**	0.587	0.675	0.883	0.441	0.262
G-mean & Jac	13.067	3.397	0.614	0.663	0.900	**0.549**	0.305
Jac & FPR	13.667	4.052	**0.629**	**0.728**	**0.910**	0.439	0.152
G-mean & WRAcc	15.800	3.420	0.562	0.674	0.781	0.240	0.103
Jac & WRAcc	14.467	3.041	0.539	0.674	0.780	0.258	0.133
SupDiff & Jac	14.533	3.502	0.605	0.692	0.902	0.443	0.210
TPR & FPR	13.633	3.485	0.602	0.687	0.899	0.438	0.207
WRAcc & SupDiff	15.600	3.793	0.568	0.696	0.832	0.217	**0.069**

The best result obtained for each quality measure analysed

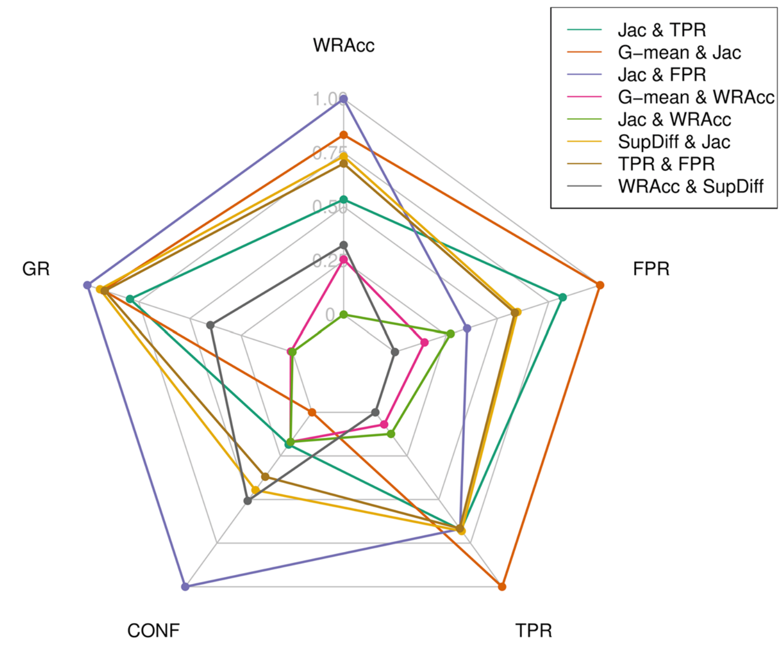

Fig. 2 Comparison of combinations. Spider chart of the different combinations of quality measures proposed in this study. The values were normalised, where a value of 1 means the maximum value extracted for this measure and 0 means the minimum one

- CONF. The combination of quality measures with the best results are Jac and FPR. As mentioned before, the generality objective is less relevant than reliability in this combination. So, rules obtained presents high levels of accuracy.
- GR. Once again, the use of Jac and FPR as objectives led to obtain the best results in this measure. Here, the focus on reliability guarantees the extraction of rules with high accuracy. This fact, together with the generality component of the Jac measure, is enough for the extraction of rules which are real EPs.
- TPR and FPR. The combination of G-mean and Jac obtain the best results in TPR. For FPR, WRACC and SuppDiff are the best. Nevertheless, it is important to remark that in this work a trade-off between generality and reliability is searched. In this way, it can be observed that the biggest difference between TPR and FPR is for Jac and FPR, which means that this combination finds the best trade-off between generality and reliability.

As can be observed, the combination of quality measures based on Jac and FPR offers the best trade-off amongst the relevant aspects of a descriptive rule in EPM. This can be produced because of FPR tries to only maximise precision whereas Jac tries to improve the generality without decreasing the precision, which produces more interesting rules.

Additionally, for each of the aspects or objectives previously mentioned for EPM, it can be observed that the combination that extracts the best results in reliability, measured as CONF, is the combination of Jac and FPR. This is due to the high reliability component of both measures. On the other hand, the combination that extracts the best results in generality, measured as TPR, is the one formed by G-mean and Jac. In this case, although G-mean has both reliability and generality components, it benefits generality over reliability. Finally, with respect to interest, measured as WRACC, there are several

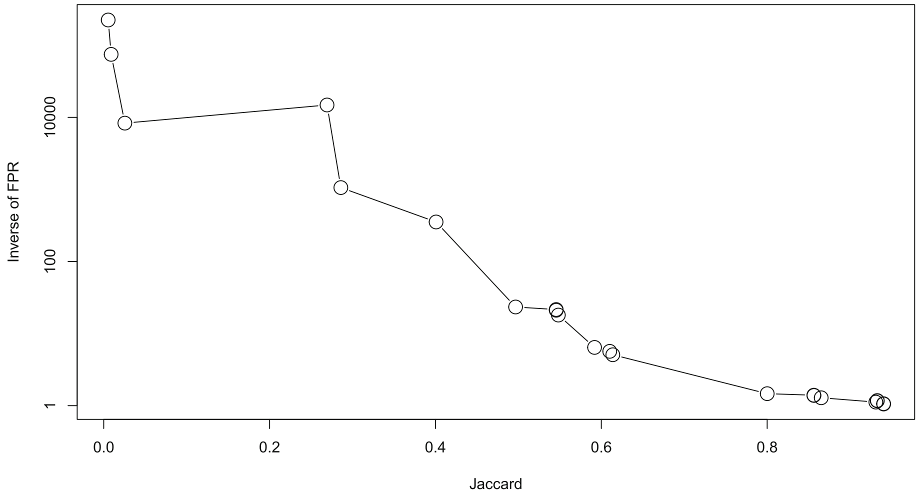

Fig. 3 Pareto front of Jaccard and FPR combination on *census* dataset Graphical representation of the Pareto front extracted on *census* dataset for Jac and FPR combination. It is important to remark that FPR was used as $\frac{1}{FPR}$ in order to minimise the FPR

combinations very close to each other which means that they have similar behaviour. However, the most interesting rules are extracted by the combination of Jac and FPR. This is due to good trade-off between reliability and generality.

As mentioned before, the rules presented to the expert are the ones that belongs to the Pareto front. In addition the rules can be presented in a chart where the Pareto front is represented. In this way, the expert is able to choose the rules that are the most important according to their objectives. In Fig. 3 the Pareto front for the *census* dataset for the Jac and FPR combination is shown. Additionally, Fig. 4 presents the pareto for the G-mean and Jac combination on *census* dataset. From these graphs it can be observed that the Pareto fronts extracted are well-suited for the extraction of high-quality descriptive rules.

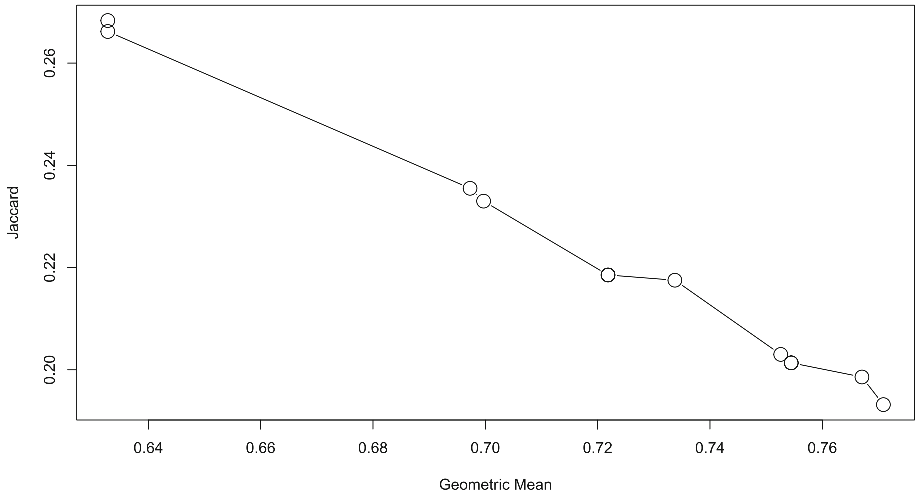

Fig. 4 Pareto front of G-mean and Jaccard combination on *census* dataset Graphical representation of the Pareto front extracted on *census* dataset for Geometric mean and Jaccard combination

Conclusions

This paper presents a study about the suitability of the combination of quality measures in order to be used as objectives for a multi-objective approach for the extraction of EPs in Big Data environments. In particular, this study presents an analysis between eight combinations of quality measures which presents some conflicts among them. The algorithm used is the BD-EFEP algorithm, a multi-objective evolutionary algorithm focused in Big Data. The study demonstrates that the selection of the objectives has a significant effect in the final result as they are used to guid the search process. In particular, the combinations formed by TPR and FPR, Geometric mean and Jaccard index, and Jaccard index and FPR are the most suitable for the relevant aspects of descriptive rules in EPM, i.e., interest, generality and reliability, respectively. Additionally, it is highlighted the performance of Jaccard index and FPR as the combination with the best trade-off in these aspects. Future works related to the extraction of EPs in Big Data are the development of new distributed approaches focused in an efficient extraction of patterns with high values in the Jaccard index and in FPR in order to improve the results extracted.

Abbreviations

Conf: Confidence; EPM: Emerging pattern mining; EPs: Emerging patterns; FPR: False positive rate; G-mean: Geometric mean of true positive rate and true negative rate; GR: Growth rate; IRL: Iterative rule learning; Jac: Jaccard index; MOEA: Multi-objective evolutionary algorithm n_r: Average number of rules; n_v: Average number of variables; OVA: One-vs-All; SupDiff: Support difference; TNR: True negative rate; TPR: True positive rate; WRAcc: Weighted relative accuracy

Acknowledgements

Not applicable.

Funding

This work was supported by the FPI scholarship BES-2016-077738, under the project TIN2015-68454-R (FEDER Founds) of the Spanish Ministry of Economy and Competitiveness.

Availability of data and materials

The BD-EFEP algorithm is available at our GitHub repository (https://github.com/SIMIDAT/bd-efep), data employed in this study come from the UCI Machine Learning repository (https://archive.ics.uci.edu/ml/index.php).

Authors' contributions

Conceived and designed the analysis: CJC, MJJ. Analyzed the data: AMGV, PG. Wrote the paper: AMGV, PG, Performed the programming: AMGV. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent to publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Computer Science, University of Jaén, Paraje Las Lagunillas, s/n, 23071 Jaén, Spain. [2]Leicester School of Pharmacy, De Montfort University, LE1 9BH Leicester, UK.

Received: 16 May 2018 Accepted: 11 December 2018
Published online: 07 January 2019

References

1. InternetLiveStats.com: Internet Live Stats. http://www.internetlivestats.com/one-second/. Accessed 05 Feb 2018.
2. Minelli M, Chambers M, Dhiraj A. Big Data, Big Analytics:Emerging Business Intelligence and Analytic Trends for Today's Businesses, 3rd edn. United States: Wiley; 2013.
3. Dean J, Ghemawat S. Mapreduce: Simplified data processing on large clusters. In: Operating Systems Design and Implementation (OSDI). New York: ACM; 2004. p. 137–50.

4. Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Commun ACM. 2008;51(1):107–13.
5. White T. Hadoop: The Definitive Guide, 4th edn. Beijing: O'Reilly; 2015.
6. Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, Franklin M, Shenker S, Stoica I. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation. Berkeley: USENIX Association; 2012.
7. Dong G, Li J. Efficient mining of emerging patterns: Discovering trends and differences. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 1999. p. 43–52.
8. García-Vico AM, Carmona CJ, Martín D, García-Borroto M, del Jesus MJ. An overview of emerging pattern mining in supervised descriptive rule discovery: Taxonomy, empirical study, trends and prospects. WIREs Data Min Knowl Disc. 2018;8(1):e1231.
9. Kralj-Novak P, Lavrac N, Webb GI. Supervised Descriptive Rule Discovery: A Unifying Survey of Constrast Set, Emerging Pateern and Subgroup Mining. J Mach Learn Res. 2009;10:377–403.
10. Lepailleur A, Poezevara G, Bureau R. Automated detection of structural alerts (chemical fragments) in (eco) toxicology. Comput Struct Biotechnol J. 2013;5(6):1–8.
11. Sherhod R, Gillet VJ, Hanser T, Judson PN, Vessey JD. Toxicological knowledge discovery by mining emerging patterns from toxicity data. J Chem Inf Model. 2013;5(S-1):9.
12. Angriyasa PW, Rustam Z, Sadewo W. Non-invasive intracranial pressure classification using strong jumping emerging patterns. In: Proc. of the 2011 International Conference on Advanced Computer Science and Information System (ICACSIS). Jakarta: IEEE; 2011. p. 377–80.
13. Yu Y, Yan K, Zhu X, Wang G. Detecting of PIU Behaviors Based on Discovered Generators and Emerging Patterns from Computer-Mediated Interaction Events. In: Proc. of the 15th International Conference on Web-Age Information Management. Cham: Springer International Publishing; 2014. p. 277–93.
14. Li G, Law R, Vu HQ, Rong J, Zhao XR. Identifying emerging hotel preferences using emerging pattern mining technique. Tour Manag. 2015;46:311–21.
15. García-Vico AM, Montes J, Aguilera J, Carmona CJ, del Jesus MJ. Analysing Concentrating Photovoltaics Technology through the use of Emerging Pattern Mining. In: Proc. of the 11th International Conference on Soft Computing Models in Industrial and Environmental Applications. San Sebastián: Springer; 2016. p. 1–8.
16. Weng C-H, Tony C-KH. Observation of sales trends by mining emerging patterns in dynamic markets. Appl Intell. 2018;48:1–15.
17. García-Vico AM, Carmona CJ, González P, del Jesus MJ. A big data approach for extracting fuzzy emerging patterns. Cognitive Computation (In press).
18. Carmona CJ, del Jesus MJ, Herrera F. A Unifying Analysis for the Supervised Descriptive Rule Discovery via the Weighted Relative Accuracy. Knowledge-Based Systems. 2018;139:89–100.
19. Dong GZ, Zhang X, Wong L, Li JY. CAEP: Classification by Aggregating Emerging Patterns. In: Proc. of the Discovery Science. LNCS, vol. 1721. Berlin: Springer; 1999. p. 30–42.
20. García-Borroto M, Loyola-González O, Martínez-Trinidad JF, Carrasco-Ochoa JA. Evaluation of quality measures for contrast patterns by using unseen objects. Expert Syst Appl. 2017;83:104–13.
21. Kloesgen W. Explora: A Multipattern and Multistrategy Discovery Assistant. In: Advances in Knowledge Discovery and Data Mining. Menlo Park: American Association for Artificial Intelligence; 1996. p. 249–71.
22. Bay SD, Pazzani MJ. Detecting group differences: Mining contrast sets. Data Min Knowl Discov. 2001;5(3):213–46.
23. Tan P-N, Kumar V, Srivastava J. Selecting the right interestingness measure for association patterns. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2002. p. 32–41.
24. Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: an overview. In: Advances in Knowledge Discovery and Data Mining. Menlo Park: AAAI/MIT Press; 1996. p. 1–34.
25. Gamberger D, Lavrac N. Expert-Guided Subgroup Discovery: Methodology and Application. J Artif Intell Res. 2002;17:501–27.
26. Kubat M, Matwin S. Addressing the curse of imbalanced training sets: One-sided selection. In: Proc. of the 14th International Conference on Machine Learning, vol. 97. Nashville: Morgan Kaufmann; 1997. p. 179–86.
27. Wang L, Zhao H, Dong G, Li J. On the complexity of finding emerging patterns. Theor Comput Sci. 2005;335(1): 15–27.
28. Dean J, Ghemawat S. MapReduce: A flexible data processing tool. Commun ACM. 2010;53(1):72–77.
29. Ramírez-Gallego S, Fernández A, García S, Chen M, Herrera F. Big data: Tutorial and guidelines on information and process fusion for analytics algorithms with mapreduce. Inf Fusion. 2018;42:51–61.
30. Peralta D, Río S, Ramíez-Gallego S, Triguero I, Beníez JM, Herrera F. Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach. Mathematical Problems in Engineering. 2015;2015:1–11.
31. Rodríguez-Fdez I, Mucientes M, Bugarín A. FRULER: fuzzy rule learning through evolution for regression. Information Sciences. 2016;354:1–18.
32. Padillo F, Luna JM, Ventura S. An evolutionary algorithm for mining rare association rules: A big data approach. In: 2017 IEEE Congress on Evolutionary Computation (CEC). San Sebastián: IEEE; 2017. p. 2007–14.
33. Padillo F, Luna JM, Herrera F, Ventura S. Mining association rules on big data through mapreduce genetic programming. Integrated Computer-Aided Engineering (In Press). 20181–19.
34. García-Vico AM, González P, del Jesus MJ, Carmona CJ. A first approach to handle emergining patterns mining on big data problems: The evaefp-spark algorithm. In: IEEE International Conference on Fuzzy Systems. Naples: IEEE; 2017. p. 1–6.
35. Cordón O, del Jesus MJ, Herrera F, Lozano M. MOGUL: A Methodology to obtain genetic fuzzy rule-based systems under the iterative rule learning approach. Internation Journal of Intelligent Systems. 1999;14:1123–53.
36. Wong ML, Leung KS. Data Mining Using Grammar Based Genetic Programming and Applications, 1st edn. Norwell: Kluwer Academics Publishers; 2000.
37. Leung KS, Leung Y, So L, Yam KF. Rule Learning in Expert Systems Using Genetic Algorithm: 1, Concepts. In: Jizuka K, editor. Proc. of the 2nd International Conference on Fuzzy Logic and Neural Networks. Japan: Fuzzy Logic Systems Institute; 1992. p. 201–204.

38. Buckland M, Gey F. The relationship between recall and precision. J Am Soc Inf Sci. 1994;45(1):12–19.
39. Ishibuchi H, Tsukamoto N, Hitotsuyanagi Y, Nojima Y. Effectiveness of scalability improvement attempts on the performance of nsga-ii for many-objective problems. In: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation (GECCO '08). New York: ACM; 2008. p. 649–56.
40. Dheeru D, Karra Taniskidou E. UCI Machine Learning Repository. 2017. http://archive.ics.uci.edu/ml.
41. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The weka data mining software: an update. ACM SIGKDD Explor Newsl. 2009;11(1):10–18.