Inf Syst Front (2018) 20:401–416
DOI 10.1007/s10796-016-9690-6

A multi-objective model for discovering high-quality knowledge based on data quality and prior knowledge

Qi Liu [1,2] · Gengzhong Feng [1,2] · Nengmin Wang [1,2] · Giri Kumar Tayi [3]

Published online: 18 August 2016

Abstract Discovering knowledge from data means finding useful patterns in data, this process has increased the opportunity and challenge for businesses in the big data era. Meanwhile, improving the quality of the discovered knowledge is important for making correct decisions in an unpredictable environment. Various models have been developed in the past; however, few used both data quality and prior knowledge to control the quality of the discovery processes and results. In this paper, a multi-objective model of knowledge discovery in databases is developed, which aids the discovery process by utilizing prior process knowledge and different measures of data quality. To illustrate the model, association rule mining is considered and formulated as a multi-objective problem that takes into account data quality measures and prior process knowledge instead of a single objective problem. Measures such as confidence, support, comprehensibility and interestingness are used. A Pareto-based integrated multi-objective Artificial Bee Colony (IMOABC) algorithm is developed to solve the problem. Using well-known and publicly available databases, experiments are carried out to compare the performance of IMOABC with NSGA-II, MOPSO and Apriori algorithms, respectively. The computational results show that IMOABC outperforms NSGA-II, MOPSO and Apriori on different measures and it could be easily customized or tailored to be in line with user requirements and still generates high-quality association rules.

Keywords Data mining · Data quality · KDD · Decision making · Multi-objective algorithm

✉ Giri Kumar Tayi
gtayi@albany.edu

Qi Liu
minkex@stu.xjtu.edu.cn

Gengzhong Feng
gzfeng@mail.xjtu.edu.cn

Nengmin Wang
wangnm@mail.xjtu.edu.cn

[1] School of Management, Xi'an JiaoTong University, NO. 28 Xianning Road, Xi'an Shaanxi 710049, China

[2] The key lab of the ministry of education for process control and efficiency engineering, NO.28 Xianning Road, Xi'an Shaanxi 710049, China

[3] School of Business, SUNY at Albany, Albany, NY 12222, USA

1 Introduction

The use of the Internet and the consequent explosive growth of information have made large volumes of a variety of data available to both businesses and individuals (M. S. Chen et al. 2006). Enterprises that could master the necessary methodology to exploit that data will improve innovation, increase competitiveness and enhance productivity (Manyika et al. 2011; Popovic et al. 2015). In the early 1990s, the process of knowledge discovery in databases (KDD) was first developed and it created the context for developing tools that could control the flood of data entering databases that are owned and operated by business, manufacturing firms, scientific organizations and personal information sources (Piatetskyshapiro 1991). Various models of KDD processes and methodologies were developed, and the original KDD process (Fayyad et al. 1996) and cross-industry standard process for data mining (CRISP-DM) (Chapman et al. 2000) are two well-known models among them. However, both of these models do not consider the significant role played by data quality and prior knowledge; thus, not all of the data mining results and

processes are useful and correct. At the same time, it is true that a large number of data-mining projects that have been developed at great expense are not considered successful because the project results are not being used (Mariscal et al. 2010). A key reason leading to the above situation is that the quality of the used data or the mined knowledge is low. Actually, data quality problem occurs in the whole KDD process. In this paper, we propose a model that emphasises the use of data quality (DQ) measures and prior knowledge to improve the quality of the results.

In order to illustrate the logic of the model, we use the Association rule mining, an important technique in data mining (DM) step of the KDD process (Han and Kamber 2006), as an example. A number of algorithms have been developed for generating the association rules (Ceglar and Roddick 2006). In traditional association rule mining, a high confidence rule will be generated, for example, the rule: "Zip code: 20,015- > City: Washington" that has a confidence of slightly below 100 %, but it may be uninteresting. However, rules at a much lower confidence level are also worth considering (Hipp et al. 2001). At the same time, the key characteristics of useful association rules are novelty, externally significant, unexpected, and actionable (Agarwal et al. 2001; Rak et al. 2008). Thus, only using *confidence* and *support* as criteria, in all likelihood, will not produce rules that are useful from a practical perspective. A generated rule which satisfies the data quality measures articulated by managers could be deemed useful. Some data quality measures such as comprehensibility, interestingness and timeliness etc. could be applicable to the generated rules (Guerra-García et al. 2013). Interestingly these measures are also highly customizable according to prior knowledge of the practical situation or decision for which the mining process is being applied. In a recent study measures were used to assess the quality of the raw data, which could improve the efficiency of the mining algorithms (Davidson and Tayi 2009). However, their paper only considers a single objective while evaluating the efficacy of classification rule mining algorithms. Lahiri and Dey (2013) used multi-objective mining algorithm to discover high quality rules, which was proved to be superior to the other proposed algorithms that only use one evaluation measure. Data quality in their paper is only used to improve the data mining algorithm. Janjua et al. (2013) developed a method to improve the quality of the integrated knowledge. However, they only considered the data quality problem in data integration process. In contrast, data quality is considered in the whole knowledge discovery process in our model. We illustrate how to use the data quality to change the traditional knowledge discovery process in Section 3. And we also propose a multi-objective Artificial Bee Colony algorithm (IMOABC), which outperforms NSGA-II, MOPSO and Apriori algorithms, for association rule mining.

The main contributions of this paper are summarised as follows:

1. A model of KDD that makes the discovery processes and the corresponding results to be more practical as it emphasises the use of data quality measures and prior knowledge in an interactive manner which truly improves the quality of the results and also reduces the complexity of the data mining task is presented.
2. To illustrate how the model could be used in reality, association rule mining is taken as an example. In the suggested KDD process, data quality measures and prior knowledge are taken as an objective and/or as a constraint, which offers an advantage over the traditional approach of association rule mining, where only *support* and *confidence* are used to extract useful rules from the raw data. Thus, the association rule mining is considered as a multi-objective problem rather than a single-objective problem, which mitigates some of the limitations of the existing approaches.
3. Subsequently, an integrated multi-objective Artificial Bee Colony algorithm is developed and tested on publicly available real data sets; the results of the experiments for both the IMOABC and Apriori algorithm are presented. The computational results show that IMOABC could be easily customized or tailored to be in line with user requirements and still generate high-quality association rules; especially the number of the generated rules is suitable for users as it does not overload their cognition capabilities. Further, different rules could be chosen according to users' different preferential weighting of the objectives using a Pareto-based approach.

This paper is organised as follows. Section 2 introduces KDD, DQ, DM and data quality mining briefly. A model for KDD that is based on data quality and prior knowledge is developed in Section 3. In Section 4, association rule mining that is based on data quality and prior knowledge is presented as an example for the model proposed in Section 3, and then, the original ABC algorithm is modified to solve the multi-objective problem. Finally, experiments are performed, and the results are discussed. Section 5 concludes the paper and proposes future research directions.

2 Background

2.1 Knowledge discovery in databases

KDD is a phrase that describes the process of discovering and exploiting the considerable amount of valid, novel, potentially useful knowledge from databases. Knowledge discovery enables information to be transformed into knowledge that is

regarded as hidden in the vast databases and can contribute to the development of knowledge innovation and a knowledge economy. It is well known that standard statistical techniques are simply not effective in discovering interesting knowledge from large size databases which are quite common in modern businesses (Corne et al. 2012).

A typical KDD project contains a series of complex mining steps. Fayyad et al. (1996) summed up the five basic steps of the KDD, as follows:

Step 1. Selection: users often should make sure what type of data could be applied to their KDD project.
Step 2. Pre-processing: when the data are collected, the next step must be pre-processing the data to eliminate errors that exist in the data and fix the missing information.
Step 3. Transformation: conversion of the data to the required format of the data mining algorithm being used. This step is critical for obtaining useful and hidden knowledge from the vast amount of raw data thereby ensuring a successful KDD project.
Step 4. Data mining: selection and application of appropriate data mining tools.
Step 5. Interpretation and evaluation: understanding and evaluating the data mining results.

Many studies have extended the KDD process in the past. Gertosio and Dussauchoy (2004) added the process with human interaction. Cabena et al. (1998) added the feedback from the result to the raw data into the original KDD process. Kros et al. (2006) presented a neural network that is helpful on KDD in the presence of imprecise data. The KDD process was also modified to be suitable for some domains (Bendoly 2003). Actually, the above models did not consider data quality measures or using prior knowledge which is important to the quality of KDD process.

2.2 Data mining

Data mining is the core of the KDD process. Using approximate algorithms to extract useful information and knowledge from large databases has been recognised by many researchers as a key research topic in database systems and machine learning and by many industrial companies as an important area that has an opportunity for major revenues (M. S. Chen et al. 1996; Liu and Shih 2005). The discovered knowledge could be applied to many areas, such as information management, marketing, decision making, and process control (Bose and Mahapatra 2001; Hui and Jha 2000; Li et al. 2007; Lin et al. 2003).

Associate rule mining is an important issue in data mining. Many of the existing algorithms, such as Apriori (Agrawal et al. 1993), SETM (Houtsma and Swami 1995), and DIC (Brin et al. 1997) for mining association rules, are mainly based on the approach that is suggested by Agrawal et al. (Agrawal et al. 1993; Agrawal and Srikant 1994). These algorithms involve two main steps:

Step 1: Frequent itemsets generation. Frequent itemsets are detected from all possible itemsets by using a measure called support count (*SUP*) and a user-defined parameter called minimum support. *SUP* is defined as

$$SUP(A \cup C) = |A \cup C| / |D| \quad (1)$$

where $|A \cup C|$ means the number of records of A and C that occur at the same time in the database, and $|D|$ means the total number of records in the database. For example, the *support* $SUP(A)$ of an itemset A is defined as the proportion of transactions in the data set which contain the itemset. That is, the itemset {tennis racket, tennis ball, sneaker} has a support of 0.2 since it occurs in 20 % of all transactions.

Step 2: Calculate the *confidence* and generate the rules using another user-defined parameter called the minimum confidence (*accuracy*). The *confidence* is defined as

$$\text{Confidence} = SUP(A \cup C) / SUP(A) \quad (2)$$

One limitation of these algorithms is that they only consider the same measures (i.e. confidence and support for mining association rule) for all the situations. This neglects the differences among the variety of situations an organization encounters and it also disregards the preferences of different managers about how the generated rules should be used. Another limitation is as follows: the number of combinations of attributes that form the rules might be very large when the database has a large number of attributes. According to the above algorithms, whose criterion relies solely on the number of occurrences of the rule in the entire database, the generated rule could have a large number of attributes, which makes it difficult for users or managers to easily understand them (Fidelis et al. 2000). If the users do not understand the meaning of the rules, then they will not use the rules. Moreover, some less interesting rules extracted could be easily predicted by the users (Freitas 2002; Noda et al. 1999).

2.3 Data quality

In this paper, data quality is taken to be synonymous to information quality. Wang and Strong (1996) developed a framework for representing the data quality dimensions, which is important to data consumers, as ascertained through a rigorous survey. They categorised the data quality into four aspects:

Intrinsic DQ denotes that data have quality in their own right. *Contextual DQ* highlights the requirement that data quality must be considered within the context of the task at hand. *Representational DQ* and *Accessibility DQ* emphasise the importance of the roles of the systems. That is, the system must be accessible but secure, and the system must present data in such a way that they are interpretable, easy to understand, and represented concisely and consistently. It is well known that poor data quality truly leads to serious and disastrous situations for any organization (Fisher and Kingma 2001). Over the past decade, data quality research activities have increased significantly to meet the needs of organisations that attempt to measure and improve the quality of the data (Madnick and Zhu 2006; Parssian et al. 2004; Pipino et al. 2002).

2.4 Data quality mining

The definition of data quality mining (DQM) is "*the deliberate application of data mining techniques for the purpose of data quality measurement and improvement. The goal of DQM is to detect, quantify, explain, and correct data quality deficiencies in very large databases*" (Hipp et al. 2001). It has been proven that it is difficult to extract the most interesting rule because of the large size of the dataset. The generated rule could have a large number of attributes, which would thereby make it difficult to understand. If the generated rules are not understandable to the user, then the user will never use them. Thus, it is important to modify the collected data and mining algorithms so as to include the subjective knowledge of the users. In this paper, the authors consider the widely used association rule mining technique and attempt to improve the practical usefulness of the technique. The authors develop an integrated multi-objective approach that uses Artificial Bee Colony (ABC) algorithm to the mining rules that are generated by explicitly incorporating two data quality objectives: *comprehensibility* and *interestingness* as well as two constraints: *confidence* and *support*.

2.5 Multi-objective algorithm

To solve the multi-objective problem, methods such as Weighted sum method, Lexicographic method and Pareto frontier could be used (Beiranvand et al. 2014). A large number of Pareto-based multi-objective heuristic algorithms have been reported in recent years, such as Elitist Non-dominated Sorting Genetic Algorithm (NSGA-II) (de la Iglesia et al. 2006), Strength Pareto Evolutionary Algorithm (SPEA2) (Zitzler et al. 2001), Multi-objective Particle Swarm Optimization (MOPSO) (Coello et al. 2004) and Modified Simulated Annealing Algorithm (MSAA) (Nasiri et al. 2010). In order to solve the data mining problem, some scholars modified above multi-objective algorithm for classification problem (B. Alatas and Akin 2009; de la Iglesia et al. 2006; Reynolds and de la Iglesia 2009), association problem (Bilal Alatas et al. 2008; Beiranvand et al. 2014; Cui et al. 2011) and feature selection (Sikora and Piramuthu 2007) etc.

Similar to the above algorithms, the Pareto-frontier approach is adopted in developing the modified algorithm in this paper. Our multi-objective algorithm is based on Artificial Bee Colony (ABC) algorithm (Karaboga 2005; Karaboga and Basturk 2008). The original ABC algorithm was developed to solve single objective problems, and it was proved to be better than Particle Swarm Optimization (PSO) and Genetic Algorithms (GA) (Karaboga and Basturk 2008); then, it was modified to solve the discrete problem (Szeto et al. 2011).

The original ABC algorithm (Karaboga 2005; Karaboga and Akay 2009; Karaboga and Basturk 2007, 2008) analyses function optimization problems through simulations of bees' foraging behaviour. It divides the bees into three bee types: employed bees, onlooker bees and scout bees. Employed bees and onlooker bees gather honey and scout bees ensure that there is not enough variation of food sources. Each employed Bee corresponds to one food source. The numbers of employed bees, onlooker bees and food sources are equal. Each food source represents a feasible solution. The quality of the food source is represented by the fitness value, which indicates the merits of the feasible solution. Employed bees communicate the fitness of the food source through a dance. Onlooker bees observe the dance of several employed bees and choose a food source from which to gather honey based on the fitness values communicated by the employed bees. Better food sources attract a larger number of onlooker bees. Employed bees abandon food sources when the fitness value is very low. When food sources are abandoned, scout bees search for a new food source. Therefore, the processes by which bees gather honey (i.e., look for high quality food sources) are similar to the processes that search for the optimal solution of a problem. The main steps of the ABC algorithm are given below:

Send the scouts onto the initial food sources.
REPEAT

(1) Send the employed bees onto the food sources and determine their nectar amounts by following behaviour:

$$V_{ij} = X_{ij} + R_{ij}*\left(X_{ij} - X_{kj}\right) \tag{3}$$

where, $i = 1, \cdots, ColonySize/2, j = 1, \cdots, Dim$, V_{ij} is the position of the new nectar source, X_{ij} is the position of source i on dimension j, X_{kj} is the position of a random source k not equal to i on dimension j and R_{ij} is a random number on the interval $[-1,1]$. R_{ij} controls the search range.

(2) Using roulette method to calculate the probability value of the sources with which they are preferred by the onlooker bees, then move the onlooker bees onto the food sources and determine their nectar amounts also by Eq. (3).
(3) Stop the exploitation process of the sources abandoned by the bees.
(4) Send the scouts into the search area for discovering new food sources, randomly.
(5) Memorize the best food source found so far.

UNTIL (requirements are met).

3 A model for KDD based on data quality and prior knowledge

One of the ten challenging problems in data mining research is to build a new methodology to help users avoid many common data mining mistakes (Yang and Wu 2006). Furthermore, standardisation of a data mining process model should be an essential research line in the present and future of data mining and knowledge discovery (Kurgan and Musilek 2006). Many models have been developed for the data mining process in the past, but a model that focuses on both data quality and prior knowledge has not been developed yet.

The original KDD process and CRISP-DM are two main models in the data mining and knowledge discovery process, while other models are mostly the variants of these two models (Mariscal et al. 2010). One of the drawbacks of the general KDD process and CRISP-DM is the lack of a measurement and improvement mechanism for data quality in the data integration step. For example, one drawback is that even if the data is cleaned in the pre-processing step, following the "garbage in garbage out" (Y. W. Lee 2006) rule, raw data still brings data quality problems. This relationship will certainly affect the quality of the data warehouse and results, such as two synonymous but different words and different meanings of the same word would pollute the data warehouse. Thus, by focusing on data quality and on prior knowledge, an interactive model of KDD that attempts to integrate subjective and objective perspectives into a unifying process-centric framework is developed. The knowledge discovery process which is based on data quality and prior knowledge (DQPK-KDD) is depicted in Fig. 1. Data quality and prior knowledge as complements to the data mining process could improve the quality of resulting information products and enhance the speed of mining. Additionally, the added feedback process brings knowledge to raw data and enhances human intuition, which could be useful to the whole process. The concept of data quality and prior knowledge also could be used to modify the traditional methods that are used in data analysis. One of the modified methods could be seen in Section 4.

3.1 Integration

In this era of big data, large amounts of data are produced every minute; however, some of the data are distributed and stored in different areas or systems, and hence frequent integration is required. However, this is the stage where data quality problems begin to surface and affect the integration process. For users, one way to mitigate the data quality problems is to use their knowledge about the firm's business environment and operations to select relevant data which could be included in the integration step. Besides, integration influences all of the processing and quality of the discovered knowledge. For this reason, this critical integration step is added although it was neglected in the original KDD process. Batista and Salgado (2007) proposed a DQ criteria analysis in data integration and showed that incorporating DQ aspects into data integration is beneficial (J. Lee and Prékopa 2013). Meanwhile prior knowledge could enable reduction in the number of resources that are required for this integration step.

3.2 Selection

Fayyad et al. (1996) indicated that massive datasets and high dimensionality is a challenge for KDD process. An additional challenge is using users' knowledge to reduce the size of the dataset selected from data warehouse, which is referred to data understanding. In the selection step, the data quality should be used as a judgment criterion to improve the quality of the data set at the beginning of selection, such as using data semantics to reduce misinterpretation (Madnick and Zhu 2006) enhancing consistency, one of data quality measures. In fact, data semantics offers several methods that could be used to improve consistency (Evangelopoulos et al. 2010). Through the application of this selection step, target data for KDD process could be prepared.

3.3 Pre-processing

In this step, data cleaning is conducted to reduce noise or outliers and it could use data quality measures as a judgment criterion for cleaning out low quality data (i.e., incomplete, unknown, timeless, or inaccurate data). Some methods for data cleaning, such as record linkages, are based on using a statistical model for determining and improving the quality of the data (Winkler 2004).

3.4 Data Mining

Data mining could be regarded as an algorithmic process that extracts patterns such as classification rules, association rules, or summaries and has also been applied to improving operations research techniques (Corne et al. 2012; Feelders et al. 2000). Recent research also shows that by explicitly considering the

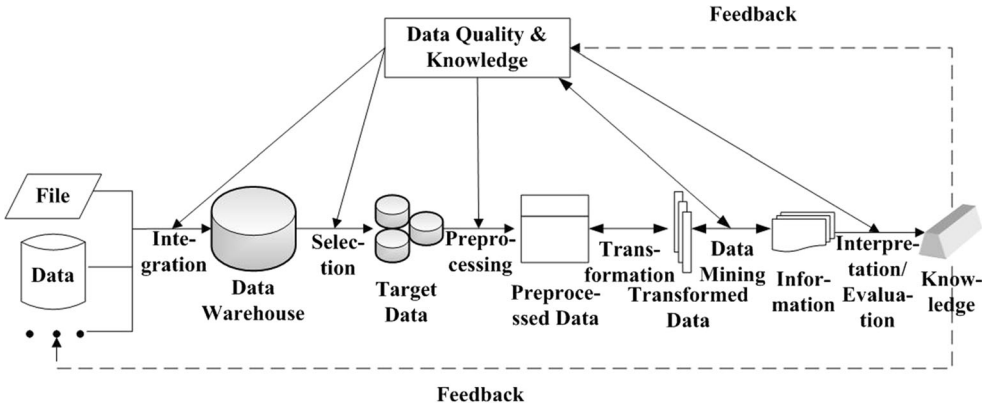

Fig. 1 The DQPK-KDD model

quality of the data that is to be mined, the performance of different mining algorithms could be significantly improved (Davidson and Tayi 2009). More and more research views classification and association rule mining process as a multi-objective process (Alhajj and Kaya 2008), which has proven to be a better approach. Application of the Apriori algorithm for mining association rules, very often generates a large number of rules (Klemettinen et al. 1994; Tan and Kumar 2000) which results in users being unnecessarily overwhelmed with incomprehensible rules. By explicitly considering data quality measures either as the objective or including them as soft constraints could yield more interesting, comprehensible and accurate rules that are useful to business users (Hipp et al. 2001).

Moreover, there are three major data quality problems in the database: incomplete data, wrong data and duplicate data. Users could use data mining technology to fill in the missing data, confirm the incorrect data and duplicate data (Wickramaratna et al. 2009) or evaluate the data quality (Soler and Yankelevich 2001) as well as give a label to the quality of the data (Sheng et al. 2008).

3.5 Evaluation

In this phase, a review is carried out to verify if the rules obtained through the KDD process achieve the stated business objectives. Next, ranking and filtering of the entire set of generated rules could be performed to evaluate and extract the subset of useful rules. Appropriate data quality measures could be used in developing the ranking (i.e. association rules could be ranked by interestingness measure which could be seen in Section 4.1), while users could use their hard and soft knowledge about the business context to facilitate the filtering of unwanted rules.

3.6 Feed-back

The knowledge gained from the mining process could be stored and archived for use in the future. Furthermore, it could be used to guide the whole process and to correct any past mistakes.

The different steps of the KDD process as outlined above could benefit immensely from actively considering the inclusion of data quality and prior knowledge. It could also help the users to control the KDD process in a meaningful way by eliminating poor quality data early in the KDD process and tailoring the mining results to be more in accordance with the requirements of the users.

4 The application of the DQPK-KDD model: using association rule mining as an example

To illustrate the model that is proposed in Section 3, an application is presented that involves association rule mining which has received considerable attention (G. Chen et al. 2006; Coenen et al. 2004). Association rule mining supports decision making extensively in business settings such as super markets, major retail stores. An association rule is an implication $A \rightarrow C$, where A is called the Antecedent, and C is called the Consequent. Many different algorithms have been developed for finding association rules (Ceglar and Roddick 2006). Association rule mining problems have been considered to be multi-objective problems, as indicated by some scholars (Ghosh and Nath 2004; Nasiri et al. 2010). The multi-objective approach has various advantages over the previous approach (e.g., it gives a set of rules instead of one rule for each run of the algorithm) and in turn provides users more useful rules (de la Iglesia et al. 2006). Moreover, users could set some practical constraints on the multi-objective problem, which enables tailoring of the mining process to the specific business context being considered. It also allows reducing the amount of data being used in the mining process and more importantly it allows the user to generate association rules that are more interesting and relevant.

Association rules could be viewed as information products (Adomavicius and Tuzhilin 1999) and hence they may also have data quality issues. When the data quality and prior knowledge are incorporated into association rule mining process, it is advantageous to utilize a multi-objective problem

rather than a single-objective problem. Since data quality has many dimensions each of which has to be evaluated appropriately, a multi-objective approach would be a good alternative. Also in many practical contexts, a user could strategically utilize some prior knowledge to control the scope of the mining process while tailoring the results of the association rule mining to the specific practical setting. For example, if the user cares only for the relationship among the attributes A, B and C, he can set the constraints such that the resulting rules contain A, B and C only. This constraint would naturally reduce the amount of data that needs to be scanned in the mining process. Such as in the real world, a sports shop which needs to know whether there is a relationship among the tennis racket, tennis ball and sneaker, could use this prior knowledge to limit the scope of loading data attributes. To solve this multi-objective association rule mining problem, a modified Artificial Bee Colony algorithm is developed. Since a single solution for a multi-objective problem is rare, the Pareto Frontier concept (Maximiano et al. 2012) is adopted in building the algorithm.

4.1 Association rule mining based on data quality and prior knowledge

Several studies have proposed interestingness measures (De Falco et al. 2002; Geng and Hamilton 2006; Klemettinen et al. 1994; Tew et al. 2014) and comprehensibility measures (Fidelis et al. 2000) for data mining. Ghosh and Nath (2004); (Lahiri and Dey 2013) proposed a multi-objective approach for mining association rules, but it has some weaknesses, for example, they consider *confidence, comprehensibility* and *interestingness* as three objectives, but several rules with *confidence* = 0, which are meaningless, could be generated thereby reducing the practical value of their approach.

In the present study, the data quality measures of *comprehensibility* and *interestingness* are used as the objectives while *confidence (accuracy)* and *support* are used as the constraint conditions for evaluating the association rules.

Comprehensibility, as an objective, is measured by the number of attributes that are involved in the rule and it attempts to quantify the understandability of the rule. It is very difficult to quantify *comprehensibility*. A careful study of an association rule inferred that if the number of conditions involved in the antecedent part is less, the rule is more comprehensible (Ghosh and Nath 2004; Qodmanan et al. 2011). It is known that the rules may be more redundant and difficult to understand when the rules contain more attributes. *Comprehensibility* is a measure that relates to the number of attributes involved in both antecedent and consequent part of the rules. If the rule has large number of attributes, users may get confused about it hence don't use it (Bilal Alatas et al. 2008). To reflect this behaviour, an expression was derived as *Comprehensibility* = N − (Number of conditions in the antecedent part), by Das and Saha (2009) who considered that this expression serves well for the classification rule generation where the number of attributes in the consequent part is always one. Since, in the association rules, the consequent part may contain more than one attribute; this expression is not suitable for the association rule mining. Thus, they proposed *comprehensibility* to be derived as

$$\text{Comprehensibility} = \log(1+|C|)/\log(1+|A\cup C|) \tag{4}$$

Here, $|C|$ and $|A \cup C|$ are the number of attributes involved in the consequent part and the total rule, respectively. However, Eq. (3) leads to a deviation, for example, if $|A_1| = 1, |C_1| = 1$, $|A_2| = 2$, and $|C_2| = 2$, then $Comprehensibility_1 = 0.6308 < Comprehensibility_2 = 0.6826$, which is inconsistent with the concept of *comprehensibility* (Ghosh and Nath 2004). Thus, Eq. 4 is revised as follows:

$$\text{Comprehensibility} = \log(1+|C|)/\log(1+|A|)*\log(1+|A\cup C|) \tag{5}$$

Interestingness measures how interesting the rule is. Since association rule mining is a part of data mining process and extracts hidden information, it should extract only those rules that are comparatively less likely to occur in the entire database. Such a surprising rule may be more interesting to the users; which again is difficult to quantify.

Interestingness measure in this paper is derived as follows:

$$\begin{aligned}\text{Interestingness} = {} & [\text{SUP}(A\cup C)/\text{SUP}(A)] \\ & \times [\text{SUP}(A\cup C)/\text{SUP}(C)] \\ & \times [1-(\text{SUP}(A\cup C)/|D|)]\end{aligned} \tag{6}$$

which follows Ghosh and Nath (2004). Here |D| is the total number of records in the database. More interestingness measures are reviewed by some scholars (Geng and Hamilton 2006; Tan and Kumar 2000).

The *Confidence* and *Support* measures are defined in Eq. (1) and Eq. (2) in Section 1, which is the same as it is used in most prior algorithms.

Moreover, users could add some other data quality measures (i.e., timelessness) according to their objective and purpose.

4.2 Rule representation

Various representation types have been used in representing the data for multi-objective problems in previous research (Srinivasan and Ramakrishnan 2011). In this paper, one representation of the rules as proposed by Ghosh and Nath (2004)

is modified, i.e., a rule is stored in arrays but is different from the classification rule representation that is proposed by Reynolds and de la Iglesia (2009). In our representation, each attribute is associated with one extra tag bit. If the first bit is 0, then the attribute next to it appears in the antecedent part, and if it is 1, then the attribute appears in the consequent part. Note that the attribute which is absent in either of these two parts will not be included in the array, which lowers the storage requirement. For example, consider the rule: A {SEX = "Female"} and C {JOB = "Stewardess"}— > D {Height > 170 cm} shown in Fig. 2.

4.3 Integrated multi-objective Artificial Bee Colony algorithm

4.3.1 Modify the original Artificial Bee Colony algorithm to suit the multi-objective problem

The original ABC is suitable for continuous variable and single objective problem. Thus, the original ABC algorithm should be modified to fit the current multi-objective problem. The original ABC algorithm is inappropriate for the multi-objective problem because the results of ABC have a weakness in terms of diversity that is some results are often the same. Thus, the NSGA-II and ABC algorithms are integrated to overcome this weakness. The integrated multi-objective Artificial Bee Colony (IMOABC) algorithm yields more diversity meanwhile it converges faster to more accurate solutions.

To use the IMOABC algorithm to solve a multi-objective problem, some modifications are necessary.

(1) Modification of the employed bees' behaviour

The original employed bees' behaviour (Eq. (3)) is to quickly search for the optimal solution, and it is suitable for the single objective problem. But when the problem is multi-objective, this behaviour will lead to a deficiency in the diversity. To solve this problem, the idea of having a cross and mutation function from the NSGA-II is adopted.

(a) Cross function

$$V(i, \text{randval1} : \text{randval2}) = V(\text{neighbour}, \text{randval1} : \text{randval2}) \quad (7)$$

$V(i, randval1 : randval2)$ is the position of the new food source. $V(neighbour, randval1 : randval2)$ is the position of the neighbouring source. *randval*1,*randval*2 are the random numbers for the attribute, and *randval*2> *randval*1.

(b) Mutation function

$$V(i, \text{randval3}) = \max(\text{randval3}) - V(i, \text{randval3}) + 1 \quad (8)$$

*randval*3 is the random number of the attribute, and max(*randval*3) is the max value of the *randval*3 attribute.

(2) Modification of onlooker bees using a tournament to select the food source:

The multi-objective tournament is adopted to select the front *FoodNumber*/2 (the parameter *FoodNumber*is the number of food sources that is set by users) number of the new food sources that are found by the employed bees instead of by the original algorithm's roulette wheel which also results in a deficiency in the diversity. Moreover, the rules that meet the given threshold of *support* and *confidence* would be selected with greater probability.

(3) Modification of onlooker bees' behaviour

Onlooker bees' behaviour is to search for a food source in a small area guided by the employed bees, whose behaviour is different from the employed bees'. But they have the same

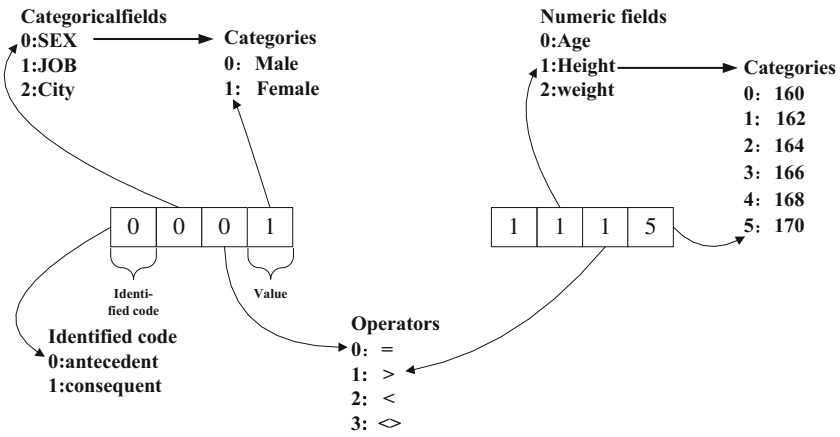

Fig. 2 The representation of an association rule

behaviour in the original ABC algorithm. Thus, its behaviour could be modified as in Eq. (9) that makes the onlooker bee in search of food near around:

$$V(i, \text{randval4}) = V(\text{neighbour}, \text{randval4}) \quad (9)$$

*randval*4 is a random number for the attribute.

(4) Modification of scout bees' behaviour

To maintain diversity in the solutions, the parameter lim*it* is set. If a source is not changed after one cycle, the tag belongs to this source which is zero at first will plus one, if the tag reach the parameter lim*it*, the food source will be saved to another archive (AR) if it belongs to F_1 (F_1 is the set of the results which could not be dominated by any other results); otherwise, it will be abandoned, and then, the corresponding employed bee becomes a scout bee, which enables it to search a new food source randomly.

(5) Adoption of a non-dominated sorting approach

To compare the results, a non-dominated sorting approach is adopted, which is imported from NSGA-II. This approach could be summarised into two parts: 1) non-dominated sort and 2) crowding distance calculate.

The above modifications to the ABC algorithm are used to make it suitable for our multi-objective problem, and the steps of the IMOABC algorithm are summarised below:

Step 1. Initialisation

Initialise the parameter lim*it*, NP, and the food source x_i which represents the association rules randomly.

Step 2. Load the records from the database.
Step 3. Scan the records to calculate the *confidence (accuracy)*, *support*, *comprehensibility*, and *interestingness* values of the rules.
Step 4. Perform fast non-dominated sorting and the fast crowded distance estimation procedure on x_i based on *comprehensibility* and *interestingness* measures.
Step 5. Select the top half of the sortedx_i.

The following steps are repeated until the conditions are met:

Step 6. Employed bees search new food sources near the original food source according to Eq. (7) or Eq. (8).
Step 7. The onlooker bees use a tournament method to select a food source, based on the information that is provided by the employed bees, and they search new food sources in the surrounding area according to Eq. (9).
Step 8. Determine whether the scout bees appear.

Repeat Steps 6–8 until the termination conditions are met or the maximum number of cycles is reached (*MaxCycles*).

Step 9. Record the food sources that belong to F_1 in AR; then, perform fast non-dominated sorting on AR.
Step 10. Output the food sources that belong to F_1 in AR.

The flow diagram of the algorithm could be seen in Fig. 3.

4.3.2 A test of the IMOABC algorithm

In order to test the performance of IMOABC, NSGA-II (Deb et al. 2002) and MOPSO (Coello et al. 2004) which are accepted as good algorithms to solve multi-objective problem by many researchers are chosen as a benchmark for comparison. IMOABC, NSGA-II and MOPSO algorithms are tested on five databases selected from UCI. The experiments have two objectives: *comprehensibility* and *interestingness.* Each algorithm runs five times, and the Pareto-based best rules obtained in each run are recorded. To evaluate the performance of the three algorithms, the average number of the rules generated is chosen as a metric. Rigorous statistical analyses were performed to test whether the new proposed algorithm has a significant improvement over the existing algorithm for given problems (Derrac et al. 2011). We adopt nonparametric statistical tests such as Wilcoxon signed rank test, Friedman test, Friedman Aligned ranks, Quade test, Contrast Estimation etc.

All experimental simulations are conducted on a computer with an AMD Athlon (tm) II X2 245 Processor, 2 GB memory, and 2.90 GHz Max Turbo Frequency. All simulations are programmed in Matlab7. For statistical analysis KEEL Software Tool (Alcalá-Fdez et al. 2009) is used. The IMOABC parameter set: the number of employed bees and the number of onlooker bees are set to be equal to the number of food sources which is half of the NP respectively are 100, 300 and 500; the parameter lim*it* is set to 0.01 * *FoodNumber* * *D* which is suggested in original algorithm, *D* is the size of the dimension of the variable; crossover and mutation probabilities are taken respectively as 0.9 and 0.1. The NSGA-II parameter set: the number of chromosome respectively is 100, 300, 500; crossover and mutation probabilities are the same as IMOABC. The MOPSO parameter set: the number of particle respectively is 100, 300, 500; the repository size is equal to the number of particle; a mutation rate of 0.5 and 30 divisions for the adaptive grid that is adopted in (Coello et al. 2004).

Table 1 shows, in detail, the results obtained by the three algorithms on five databases. Note that the number of rules generated by IMOABC, NSGA-II and MOPSO with the same number of population and same iterations is compared. Moreover, the number of generated rules is the sum of all

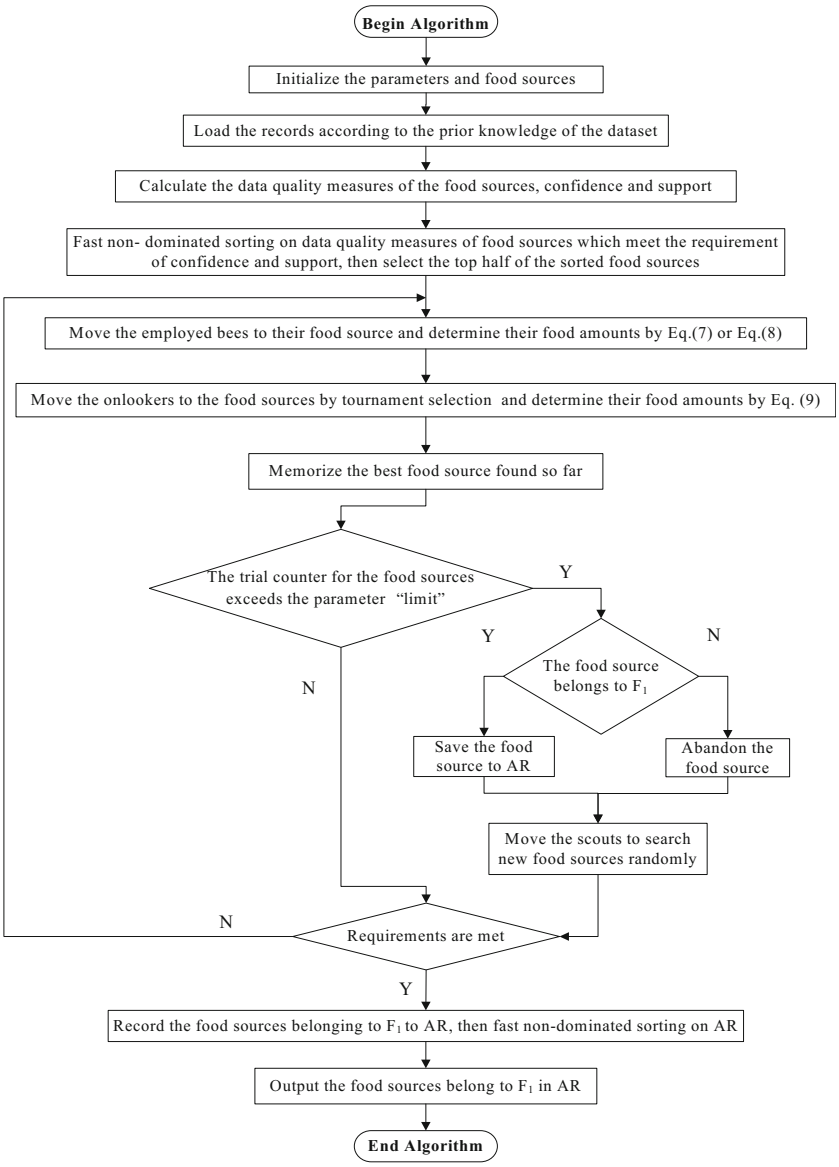

Fig. 3 Integrated multi-objective Artificial Bee Colony (IMOABC) algorithm

those rules found in the first non-dominated fronts in the set of final results obtained by IMOABC, NSGA-II and MOPSO, respectively. If there are same rules in the results, only one of them will appear in the first non-dominated front and be counted in the calculation of the number of the generated rules. For example, in the case of GCD dataset there are 12 association rules (for 100 population and 500 iteration) obtained by IMOABC while 3 association rules by NSGA-II and 0 association rule by MOPSO under the same conditions.

Through the experiments, it could be seen that IMOABC is appropriate and could be used in multi-objective association rule mining. Thus, the original Artificial Bee Colony algorithm is extended to a broader field. It could be seen from the Table 1 that the number of rule sets obtained by IMOABC is consistently more than that found by NSGA-II and MOPSO, especially in the database named Handwritten Digits which contains a larger number of attributes, for example in the case of "Population = 500, Iteration = 1500" in Handwritten Digits, 93 association rules are obtained by IMOABC while only 21 association rules by NSGA-II and 1 rules by MOPSO.

Statistical analysis is also done to evaluate the performance of the algorithms. We choose the results obtained by each algorithm in 1000 iteration for statistical analysis. Table 2 shows the R^+(where control algorithm performed better than comparing algorithm),R^-(where control algorithm performed worse than comparing algorithm), and *p*-values computed for all the pair wise comparisons concerning IMOABC as control algorithm.

The Wilcoxon test results of applying the algorithms on five databases are shown in Table 2. From the Table 2, it is clear that IMOABC shows a significant improvement over NSGA-II, MOPSO with a level of significance$\alpha = 0.05$in all dimensions.

Table 1 Comparison of performance on five databases

Dataset	Population	Iteration	Number of rules generated		
			IMOABC	NSGA-II	MOPSO
German Credit Data (GCD)	100	500	12	3	0
		1000	10	4	0
		1500	8	6	0
	300	500	21	1	0
		1000	24	5	0
		1500	5	2	0
	500	500	17	9	0
		1000	11	8	0
		1500	6	2	0
Handwritten Digits (HD)	100	500	10	9	1
		1000	16	2	0
		1500	22	2	0
	300	500	35	30	1
		1000	28	24	0
		1500	153	62	0
	500	500	63	31	1
		1000	132	48	0
		1500	93	21	1
Solar Flare Data (SFD)	100	500	5	0	0
		1000	2	2	0
		1500	10	3	0
	300	500	38	10	0
		1000	13	11	0
		1500	5	0	0
	500	500	12	0	0
		1000	2	0	0
		1500	19	0	0
Pittsburgh Bridges (PB)	100	500	17	17	1
		1000	2	0	0
		1500	17	7	1
	300	500	80	7	4
		1000	5	0	0
		1500	34	33	1
	500	500	13	0	0
		1000	6	0	0
		1500	4	0	0
Dermatology Database (DD)	100	500	13	7	1
		1000	20	5	0
		1500	14	9	0
	300	500	34	24	1
		1000	38	11	1
		1500	36	12	1
	500	500	43	22	1
		1000	47	42	0
		1500	264	0	0

The ranks of the Friedman, Friedman Aligned, and Quade tests for 50,100,500 dimensions are shown in Table 3. It implies that IMOABC gets the lowest ranks in all dimensions, highlighting IMOABC as the best performing algorithm of the comparison. The p-values computed by Friedman and Quade tests suggest that the existence of significant differences among the algorithms considered.

In order to estimate the difference between the performance of each two algorithms, Contrast Estimation is carried out and based on the medians of samples of results considering all pair wise comparison. Table 4 shows the estimations computed for each algorithm. According to the rows of the table, we might highlight the good performance of IMOABC (all its related estimators are negative which means that it achieves the lowest error rates considering median estimators). And we could also find that NSGA-II is better than MOPSO in all experiments.

The reason why IMOABC can obtain better results can be explained by examining the onlooker bees' and scout bees' behaviors. The onlooker bees help find the best food source within the limits of areas searched by employed bees, this behavior increases the probability of getting a better solution. The scout bees help find the new food sources which is different from the old food sources, this behavior enhances the ability of algorithm to jump out of local optimal solution. And the optimal results are stored in the independent storage space AR which prevents from abandoning good results. With the help of onlooker bees, scout bees and independent storage, the rules generated are not merely of higher quality but are also of greater variety.

4.4 Implementation and results

4.4.1 Datasets for experimentation

The databases that are used for this experiment are extracted from the UCI repository (Alpaydin and Kaynak 1998; Hofmann 1994). Table 5 represents their main characteristics, including the number of classes and the number of numerical and categorical attributes. Two databases are taken for experimentation that only contain attributes that are integer and categorical. In fact, there are some methods that transform numerical values to categorical values, such as using some ranges of values to modify the numerical value to the categorical.

Attribute 5, which is a real value from German Credit Data (GCD), is removed, because the attribute has only a zero value. The removal of this attribute has only a small effect on the performance of the data mining algorithm.

4.4.2 Results

Both the IMOABC and Apriori algorithms are tested on the above datasets. The IMOABC parameter is set almost the same as Section 4.3.2 except the number of food sources

Table 2 Wilcoxon signed ranks test results

IMOABC vs	Dimension								
	100			300			500		
	R^+	R^-	p-value	R^+	R^-	p-value	R^+	R^-	p-value
NSGA-II	10.0	0.0	0.04461	15.0	0.0	0.030971	15.0	0.0	0.030971
MOPSO	15.0	0.0	0.025568	15.0	0.0	0.030971	15.0	0.0	0.030971

which is set to 300 and still equal to the number of employed bees and the number of onlooker bees. The algorithm runs ten times, each run contains 1000 iterations, and the average number of the Pareto-based best rules obtained in ten runs is recorded. The support threshold is set to 10 % which is often adopted by many researchers (Gray and Orlowska 1998; Lui and Chung 2000). Further an additional support threshold level of 5 % is also considered in this experiment to facilitate comparison. Meanwhile, the threshold levels of support (5 % and 10 %) and confidence (70 % – 100 %) are set in Apriori algorithm.

4.4.3 Discussion of the results

A review of the experiment results indicates that IMOABC algorithm could be used for solving the multi-objective association rule mining problem. The original Artificial Bee Colony algorithm is extended to a broader field. The running times of IMOABC algorithm are bounded by $O(MN^2)$, where M is the number of objectives, and N is the population size which is equal to the number of the employed bees. It could be operated on most modern personal computers.

From Table 6, when the Apriori algorithm is applied to either GCD or SFD datasets it generates a large number of rules for all threshold levels of support (5 % and 10 %) and confidence (70 %– 100 %). For example, in the case of GCD dataset there are 161,172 association rules (for 5 % support and 70 % confidence threshold levels). Although the number of rules generated decreases significantly as the confidence threshold increases there are still a large number of rules, about 5633 rules at the confidence threshold level of 100 %. The same trend is evident even when the support threshold increases to 10 %. In the case of 100 % confidence level, the number of rules generated at the 5 % support level is drastically larger than that at the 10 % support level (5633 versus 211 rules). Clearly generating a large number of rules alone does not enhance the practical value of the data mining algorithm. In order to reduce the number of generated rules, users may opt to increase the threshold level of support and/or confidence (as could be seen from Table 4). However, choosing the threshold levels merely to reduce the number of rules generated could be deemed myopic and at best arbitrary and subjective. Importantly, some of the rules that may be lost in the process could potentially be of high practical value and could also be very relevant to the specific business context. As a case in point when the confidence threshold is fixed at 70 %, an increase in the support threshold from 5 % to 10 %, results in a dramatic drop (from 161,172 to 29,160) in the number of rules generated. Further not all rules are interesting just because they have high support and confidence levels. So users could add some data quality measures which could reflect the users' practical situation into association rules mining. Actually more interesting rules may be the ones with low data quality metrics such as incomprehensibility, not timely etc. To address this practical but important problem, a multi-objective problem formulation that explicitly accounts for data quality measures and prior knowledge is more effective than a single objective problem formulation. The Pareto-based results obtained by IMOABC algorithm are also presented in Table 6. Interestingly, when confidence threshold is fixed and the

Table 3 Ranks achieved by the Friedman, Friedman Aligned, and Quade tests on mean function

Algorithms	Dimension								
	100			300			500		
	Friedman	Friedman Aligned	Quade	Friedman	Friedman Aligned	Quade	Friedman	Friedman Aligned	Quade
IMOABC	1.1	3.1	1.05	1	3.4	1	1	3.8	1
NSGA-II	2	9	2	2.1	8.1	2.0333	2.2	8.2	2.1
MOPSO	2.9	11.9	2.95	2.9	12.5	2.9667	2.8	12	2.9
Statistic	8.1	3.37	11.29	9.1	3.46	15.21	8.4	3.43	11.65
p-value	1.7e-2	1.85e-1	4.68e-3	1.1e-2	1.77e-1	1.88e-3	1.5e-2	1.79e-1	4.25e-3

Table 4 Contrast Estimation results

Dimension		IMOABC	NSGA-II	MOPSO
100	IMOABC	0	6.667	9.333
	NSGA-II	-6.667	0	2.667
	MOPSO	-9.333	-2.667	0
300	IMOABC	0	8	21
	NSGA-II	-8	0	13
	MOPSO	-21	-13	0
500	IMOABC	0	4.333	11.667
	NSGA-II	-4.333	0	7.333
	MOPSO	-11.667	-7.333	0

support threshold level increases from 5 % to 10 %, changes in the number of the rules in majority of the cases is flat or unchanged thereby signalling that the results obtained by IMOABC are more robust than those obtained by applying Apriori algorithm. Because there are only two criteria (*confidence* and *support*) in Apriori algorithm it results in low robustness, when multi-objective data quality measures are added into IMOABC algorithm the number of rules generated do not change dramatically under different *confidence* and *support* threshold levels.

A set of Pareto-based rules generated by IMOABC contains different weight values of *comprehensibility* and *interestingness* (Kim and De Weck 2005) which could be customised by users. For example, in dataset GCD, there are 7 rules generated by IMOABC (for 10 % support and 70 % confidence threshold levels). The rules are shown in Fig.4**.** If the weight value of *comprehensibility* is high, users might choose rule 1 rather than rule 7.

Appropriate number of populations and iterations could get more in line with user requirements and generates high-quality association rule, especially the number of the rules is suitable for users and do not changes much at different support and confidence threshold levels.

5 Conclusions and future work

The main contribution of this paper is to give a new perspective on the KDD process, which is, in fact, based on data quality and prior knowledge. A new model, which is based on the data quality and prior knowledge, conducting the data mining process, is developed for discovering high-quality knowledge from data; the results are high quality. In other words, data quality and knowledge could improve the quality of data mining technology, and data mining could also be used to measure and improve the data quality of a data set. Then, the model is discussed how it could be used in data analysis and business decision making. By introducing DQPK-KDD, we hope to stimulate research that considers our point of view on the interaction between the data quality, prior knowledge and KDD, which has a large potential and practical significance.

Table 5 Description of datasets

Name	Records	Attribute	Numerical	Categorical
German Credit Data (GCD)	2000	19	6	13
Solar Flare Data (SFD)	1389	13	0	13

Table 6 Comparison of the Algorithms' Performance

Dataset	Support threshold (%)	Confidence (%)	Number of rules generated	
			IMOABC	Apriori
German Credit Data (GCD)	5	70	7	161,172
		80	7	110,182
		90	5	55,220
		100	6	5633
	10	70	7	29,160
		80	8	20,373
		90	7	9681
		100	5	211
Solar Flare Data (SFD)	5	70	8	14,344
		80	6	9938
		90	6	8132
		100	5	7696
	10	70	6	7292
		80	6	5290
		90	4	3736
		100	7	3736

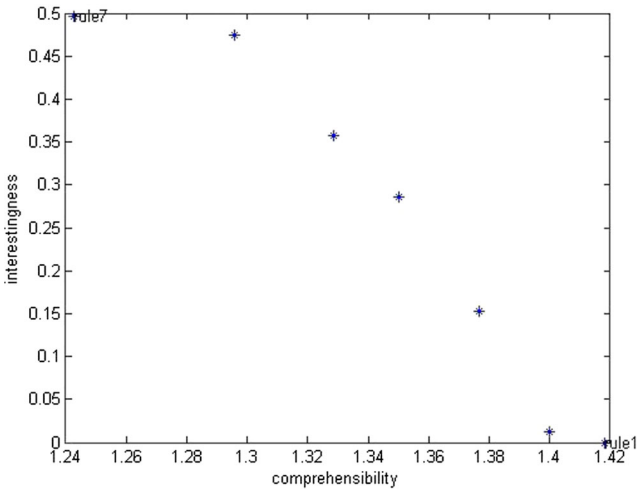

Fig. 4 The rules generated by IMOABC in dataset GCD at 10 % support and 70 % confidence

To illustrate the model, association rule mining is presented as an example. Association rule mining is viewed as a multi-objective problem rather than a single-objective problem in this paper; however, a *subjective* single objective (*support* and *confidence*) is assumed by most of the existing algorithms. The multi-objective association rule mining constrains the exploration, through the incorporation of measures of data quality which could be set by users' prior knowledge of their practical situation, then makes the rule more customizable. An integrated multi-objective Artificial Bee Colony algorithm is developed to solve the multi-objective problem; it uses two new data quality measures, *Comprehensibility* and *Interestingness,* combining with two original measures, *Confidence* (accuracy) and *Support*. The Artificial Bee Colony algorithm is also shown how it could be modified to solve a multi-objective problem, which is another contribution. The results of experiments of various well-known databases for IMOABC, NSGA-II and MOPSO algorithms are presented. The computational results show that IMOABC outperforms NSGA-II and MOPSO. Through the example of a multi-objective problem, association rule mining illustrates that the KDD that is based on the data quality and the prior knowledge is useful in reality and could be expanded to other processes.

In this paper, we only have focused on the popular association rule mining but believe that data quality and prior knowledge could be used for other tasks in KDD process such as clustering, classification rules and even reducing misinterpretation. As a trend of further research, more domains could be researched with respect to how to use the data quality and prior knowledge to enhance technologies that are used for business, data analysis, and predicting the future.

In this paper, the association rule mining only considers four measures, which could be expanded to more measures, such as timelessness, or which could be modified according to the usage. To improve the operational efficiency of the IMOABC algorithm, some methods could be used, e.g., a sample of the original database (Busygin et al. 2008), or combining a method that does not scan the whole database. Moreover, we tested the algorithms only on the databases with the categorical and integer attributes; however, the real values of the attributes could also be considered in the future research.

Acknowledgments The research presented in this paper is supported by the National Natural Science Foundation Project of China (71390333 & 71572145), the National Social Science Foundation Project of China (12&ZD070), Supported by Program for New Century Excellent Talents in University (NCET-13-0460), and the Fundamental Research Funds for the Central Universities.

References

Adomavicius, G., & Tuzhilin, A. (1999). User profiling in personalization applications through rule discovery and validation. Paper presented at the Proceedings of the fifth *ACM SIGKDD international conference on Knowledge discovery and data* mining.

Agarwal, R. C., Aggarwal, C. C., & Prasad, V. V. V. (2001). A tree projection algorithm for generation of frequent item sets. *Journal of Parallel and Distributed Computing, 61*(3), 350–371. doi:10.1006/jpdc.2000.1693.

Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*, *Paper presented at the Proc. 20th Int*. VLDB: Conf. Very Large Data Bases.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. Paper presented at the ACM SIGMOD Record.

Alatas, B., & Akin, E. (2009). Multi-objective rule mining using a chaotic particle swarm optimization algorithm. *Knowledge-Based Systems, 22*(6), 455–460. doi:10.1016/j.knosys.2009.06.004.

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.

Alcalá-Fdez, J., Sánchez, L., García, S., del Jesús, M. J., Ventura, S., Garrell, J., et al. (2009). KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing, 13*(3), 307–318.

Alhajj, R., & Kaya, M. (2008). Multi-objective genetic algorithms based automated clustering for fuzzy association rules mining. *Journal of Intelligent Information Systems, 31*(3), 243–264. doi:10.1007/s10844-007-0044-1.

Alpaydin, E., & Kaynak, C. (1998). Optical Recognition of Handwritten Digits Data Set UCI repository of machine learning databases. Retrieved from http://www.cs.uci.edu/~mlearn/MLRepository.html

Batista, M. D. C. M., & Salgado, A. C. (2007). Information Quality Measurement in Data Integration Schemas. Paper presented at the QDB.

Beiranvand, V., Mobasher-Kashani, M., & Abu Bakar, A. (2014). Multi-objective PSO algorithm for mining numerical association rules without a priori discretization. *Expert Systems with Applications, 41*(9), 4259–4273.

Bendoly, E. (2003). Theory and support for process frameworks of knowledge discovery and data mining from ERP systems. *Information Management, 40*(7), 639–647.

Bose, I., & Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. *Information Management, 39*(3), 211–225.

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. Paper presented at the ACM SIGMOD Record.

Busygin, S., Prokopyev, O., & Pardalos, P. M. (2008). Biclustering in data mining. *Computers & Operations Research, 35*(9), 2964–2987.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A., I. B. M. C., &., I. T. S. O. (1998). *Discovering data mining: from concept to implementation* (Vol. 1): Prentice Hall Upper Saddle River, NJ.

Ceglar, A., & Roddick, J. F. (2006). Association mining. *ACM Computing Surveys, 38*(2). doi:10.1145/1132956/1132958.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Chen, M. S., Han, J. W., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866–883.

Chen, G., Liu, H., Yu, L., Wei, Q., & Zhang, X. (2006). A new approach to classification based on association rule mining. *Decision Support Systems, 42*(2), 674–689.

Coello, C. A. C., Pulido, G. T., & Lechuga, M. S. (2004). Handling multiple objectives with particle swarm optimization. *Evolutionary Computation, IEEE Transactions on, 8*(3), 256–279.

Coenen, F., Leng, P., & Ahmed, S. (2004). Data structure for association rule mining: T-trees and P-trees. *IEEE Transactions on Knowledge and Data Engineering, 16*(6), 774–778.

Corne, D., Dhaenens, C., & Jourdan, L. (2012). Synergies between operations research and data mining: The emerging use of multi-

objective approaches. *European Journal of Operational Research, 221*(3), 469–479. doi:10.1016/j.ejor.2012.03.039.

Cui, J., Li, Q., & Yang, L.-P. (2011). Fast Algorithm for Mining Association Rules Based on Vertically Distributed Data in Large Dense Databases. *Computer Science, 38*(4), 216.

Das, S., & Saha, B. (2009). Data Quality Mining using Genetic Algorithm. *International Journal of Computer Science and Security, 3*(2), 105–112.

Davidson, I., & Tayi, G. (2009). Data preparation using data quality matrices for classification mining. *European Journal of Operational Research, 197*(2), 764–772.

De Falco, I., Della Cioppa, A., & Tarantino, E. (2002). Discovering interesting classification rules with genetic programming. *Applied Soft Computing, 1*(4), 257–269.

de la Iglesia, B., Richards, G., Philpott, M. S., & Rayward-Smith, V. J. (2006). The application and effectiveness of a multi-objective metaheuristic algorithm for partial classification. *European Journal of Operational Research, 169*(3), 898–917. doi:10.1016/j.ejor.2004.08.025.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation, 6*(2), 182–197. doi:10.1109/4235.996017.

Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation, 1*(1), 3–18.

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2010). Latent Semantic Analysis: five methodological recommendations. *European Journal of Information Systems, 21*(1), 70–86. doi:10.1057/ejis.2010.61.

Fayyad, U., PiatetskyShapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27–34. doi:10.1145/240455.240464.

Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information Management, 37*(5), 271–281.

Fidelis, M. V., Lopes, H., & Freitas, A. (2000). *Discovering comprehensible classification rules with a genetic algorithm.* Paper presented at the Evolutionary Computation, 2000. Proceedings of the 2000 Congress on.

Fisher, C. W., & Kingma, B. R. (2001). Criticality of data quality as exemplified in two disasters. *Information Management, 39*(2), 109–116. doi:10.1016/S0378-7206(01)00083-0.

Freitas, A. A. (2002). *Data mining and knowledge discovery with evolutionary algorithms*: Springer.

Geng, L. Q., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. ACM Computing Surveys, 38(3). doi 10.1145/1132960.1132963

Gertosio, C., & Dussauchoy, A. (2004). Knowledge discovery from industrial databases. *Journal of Intelligent Manufacturing, 15*(1), 29–37.

Ghosh, A., & Nath, B. (2004). Multi-objective rule mining using genetic algorithms. *Information Sciences, 163*(1), 123–133.

Gray, B., & Orlowska, M. E. (1998). CCAIIA: Clustering categorical attributes into interesting association rules. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 132–143). Germany: Springer Berlin Heidelberg.

Guerra-García, C., Caballero, I., & Piattini, M. (2013). Capturing data quality requirements for web applications by means of DQ_WebRE. *Information Systems Frontiers, 15*(3), 433–445.

Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques*. San Francisco: Morgan Kaufmann Publishers Inc.

Hipp, J., Guntzer, U., & Grimmer, U. (2001). *Data quality mining-making a virtue of necessity.* Paper presented at the Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD, Santa Barbara, CA, http://www.cs.cornell.edu/johannes/papers/dmkd2001-papers/p5_hipp.pdf.

Hofmann, H. (1994). Statlog (German Credit Data) Data Set UCI repository of machine learning databases. Retrieved from http://www.cs.uci.edu/~mlearn/MLRepository.html

Houtsma, M., & Swami, A. (1995). *Set-oriented mining for association rules in relational databases.* Paper presented at the Data Engineering, 1995. Proceedings of the Eleventh International Conference on.

Hui, S. C., & Jha, G. (2000). Data mining for customer service support. *Information Management, 38*(1), 1–13.

Janjua, N. K., Hussain, F. K., & Hussain, O. K. (2013). Semantic information and knowledge integration through argumentative reasoning to support intelligent decision making. *Information Systems Frontiers, 15*(2), 167–192.

Karaboga, D. (2005). An idea based on honey bee swarm for numerical optimization. *Techn. Rep. TR06, Erciyes Univ.* Press, Erciyes.

Karaboga, D., & Akay, B. (2009). A comparative study of Artificial Bee Colony algorithm. *Applied Mathematics and Computation, 214*(1), 108–132. doi:10.1016/j.amc.2009.03.090.

Karaboga, D., & Basturk, B. (2007). Artificial Bee Colony (ABC) optimization algorithm for solving constrained optimization problems. *Foundations of Fuzzy Logic and Soft Computing, Proceedings, 4529*, 789–798.

Karaboga, D., & Basturk, B. (2008). On the performance of artificial bee colony (ABC) algorithm. *Applied Soft Computing, 8*(1), 687–697. doi:10.1016/j.asoc.2007.05.007.

Kim, I. Y., & De Weck, O. (2005). Adaptive weighted-sum method for bi-objective optimization: Pareto front generation. *Structural and Multidisciplinary Optimization, 29*(2), 149–158.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A. I. (1994). Finding interesting rules from large sets of discovered association rules. Paper presented at the Proceedings of the third international conference on Information and knowledge management.

Kros, J. F., Lin, M., & Brown, M. L. (2006). Effects of the neural network s-Sigmoid function on KDD in the presence of imprecise data. *Computers & Operations Research, 33*(11), 3136–3149.

Kurgan, L. A., & Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *Knowledge Engineering Review, 21*(1), 1–24. doi:10.1017/S0269888906000738.

Lahiri, A., & Dey, D. (2013). Effects of piracy on quality of information goods. *Management Science, 59*(1), 245–264.

Lee, Y. W. (2006). *Journey to data quality*. Cambridge: MIT Press.

Lee, J., & Prékopa, A. (2013). Properties and calculation of multivariate risk measures: MVaR and MCVaR. *Annals of Operations Research, 211*(1), 225–254.

Li, T., Ruan, D., Geert, W., Song, J., & Xu, Y. (2007). A rough sets based characteristic relation approach for dynamic attribute generalization in data mining. *Knowledge-Based Systems, 20*(5), 485–494.

Lin, Q.-Y., Chen, Y.-L., Chen, J.-S., & Chen, Y.-C. (2003). Mining inter-organizational retailing knowledge for an alliance formed by competitive firms. *Information Management, 40*(5), 431–442.

Liu, D.-R., & Shih, Y.-Y. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information Management, 42*(3), 387–400.

Lui, C.-L., & Chung, F.-L. (2000). Discovery of generalized association rules with multiple minimum supports *Principles of Data Mining and Knowledge Discovery* (pp. 510-515): Springer.

Madnick, S., & Zhu, H. (2006). Improving data quality through effective use of data semantics. *Data & Knowledge Engineering, 59*(2), 460–475. doi:10.1016/j.datak.2005.10.001.

Manyika, J., Institute, M. G., Chui, M., Brown, B., Bughin, J., Dobbs, R., Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*: McKinsey Global Institute.

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and

methodologies. *Knowledge Engineering Review, 25*(2), 137–166. doi:10.1017/S0269888910000032.

Maximiano, M. D., Vega-Rodriguez, M. A., Gomez-Pulido, J. A., & Sanchez-Perez, J. M. (2012). Multiobjective metaheuristics for frequency assignment problem in mobile networks with large-scale real-world instances. *Engineering Computations, 29*(1–2), 144–172. doi:10.1108/02644401211206034.

Nasiri, M., Taghavi, L. S., & Minaee, B. (2010). Multi-Objective Rule Mining using Simulated Annealing Algorithm. *Journal of Convergence Information Technology, 5*(1), 60–68.

Noda, E., Freitas, A. A., & Lopes, H. S. (1999). *Discovering interesting prediction rules with a genetic algorithm.* Paper presented at the Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on.

Parssian, A., Sarkar, S., & Jacob, V. S. (2004). Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. *Management Science, 50*(7), 967–982. doi:10.1287/mnsc.1040.0237.

Piatetskyshapiro, G. (1991). Knowledge Discovery in Databases. *Ieee Expert-Intelligent Systems & Their Applications, 6*(5), 74–76.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211. doi:10.1145/505248.506010.

Popovic, T., Kezunovic, M., & Krstajic, B. (2015). Smart grid data analytics for digital protective relay event recordings. *Information Systems Frontiers, 17*(3), 591–600.

Qodmanan, H. R., Nasiri, M., & Minaei-Bidgoli, B. (2011). Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. *Expert Systems with Applications, 38*(1), 288–298.

Rak, R., Kurgan, L., & Reformat, M. (2008). A tree-projection-based algorithm for multi-label recurrent-item associative-classification rule generation. *Data & Knowledge Engineering, 64*(1), 171–197. doi:10.1016/j.datak.2007.05.006.

Reynolds, A. P., & de la Iglesia, B. (2009). A multi-objective GRASP for partial classification. *Soft Computing, 13*(3), 227–243. doi:10.1007/s00500-008-0320-1.

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). *Get another label? improving data quality and data mining using multiple, noisy labelers.* Paper presented at the Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.

Sikora, R., & Piramuthu, S. (2007). Framework for efficient feature selection in genetic algorithm based data mining. *European Journal of Operational Research, 180*(2), 723–737. doi:10.1016/j.ejor.2006.02.040.

Soler, S. V., & Yankelevich, D. (2001). Quality Mining: A Data Mining Based Method for Data Quality Evaluation. Paper presented at the Processing of the Sixth international Conference on Data Quality, MIT.

Srinivasan, S., & Ramakrishnan, S. (2011). Evolutionary multi objective optimization for rule mining: a review. *Artificial Intelligence Review, 36*(3), 205–248. doi:10.1007/s10462-011-9212-3.

Szeto, W., Wu, Y., & Ho, S. C. (2011). An artificial bee colony algorithm for the capacitated vehicle routing problem. *European Journal of Operational Research, 215*(1), 126–135.

Tan, P.-N., & Kumar, V. (2000). Interestingness measures for association patterns: A perspective. Paper presented at the Proc. of Workshop on Postprocessing in Machine Learning and Data Mining.

Tew, C., Giraud-Carrier, C., Tanner, K., & Burton, S. (2014). Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Mining and Knowledge Discovery, 28*(4), 1004–1045.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 5–33.

Wickramaratna, K., Kubat, M., & Premaratne, K. (2009). Predicting Missing Items in Shopping Carts. *IEEE Transactions on Knowledge and Data Engineering, 21*(7), 985–998. doi:10.1109/Tkde.2008.229.

Winkler, W. E. (2004). Methods for evaluating and creating data quality. *Information Systems, 29*(7), 531–550.

Yang, Q., & Wu, X. D. (2006). 10 Challenging problems in data mining research. *International Journal of Information Technology and Decision Making, 5*(4), 597–604. doi:10.1142/S0219622006002258.

Zitzler, E., Laumanns, M., & Thiele, L. (2001). SPEA2: Improving the strength Pareto evolutionary algorithm: Eidgenössische Technische Hochschule Zürich (ETH), Institut für Technische Informatik und Kommunikationsnetze (TIK).

Qi Liu is Ph.D. student in the Dept. of Information Management and E-Business, at the School of Management at Xi'an Jiaotong University of China. His research interests include information system management, big data and information quality and related topics.

Gengzhong Feng is Professor of Information Management and E-Business at the School of Management, Xi'an Jiaotong University, P. R. of China. He obtained the B.S. degree in computer science in 1987, the M.S. degree in systems engineering in 1990, and the Ph.D. degree in management engineering in 1993, all from Xi'an Jiaotong University of China. His research interests include logistics and supply chain management, information system management, big data and information quality. His research has been published in such journals as *International Journal of Production Research, European Journal of Operational Research, Omega and Expert Systems with Applications*.

Nengmin Wang is Professor of Industrial Engineering Department of Management School, Xi'an Jiaotong University, Xi'an, China. He received the B.S. degree in investment economy from Central South University of Technology, Changsha, China, in 1997, the M.S. degree in management science and engineering from Central South University of Technology, Changsha, China, in 2000, and the Ph.D. degree in management science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2003. His research interests include supply chain management and big data. He has authored 19 articles in refereed academic journals, including *Journal of Management Information Systems, IEEE Transactions on Engineering Management, European Journal of Operational Research, Annals of Operations Research, International Journal of Production Research, Computers & Operations Research, Computers and Industrial Engineering*.

Giri Kumar Tayi is a Professor of Management Science and Information Systems at the State University of New York at Albany. He obtained his Ph.D. from Carnegie Mellon University. His current research streams include E-Commerce and Marketing, Information Sharing in Supply Chains, Economics of Information Systems, Geographically Distributed Software Development, Data Quality and Data Mining, Open Data Initiatives and Digital Government. He has published over 55-refereed journal articles, has over 60 conference proceedings/ presentations. Many of the articles appear in top-tier academic journals such as *Operations Research, Information Systems Research, Management Science, MIS Quarterly, IEEE Transactions, Networks, Naval Research Logistics, EJOR, Journal of Combinatorial Optimization, INFORMS Journal of Computing, Journal of Computer Security, Quantitative Marketing and Economics, Government Information Quarterly, Communications of the ACM*.