Abstract— According to Scopus, only in the year 2016, there were 15747 scientific papers about data mining and KDD. These have been and remain useful technologies. In this paper, we determine the approximation level of SIA to KDD and Data Mining. To this end, we have created an approximation framework based on definition and step process proposed by Fayyad. We use mapping review and systematic review from literature published in the last 5 years in bibliographic databases ACM, EBSCO, Google Scholar, IEEE, ProQuest, Scopus and WOS. We started with 200 papers and finally, 35 had all quality criteria. This paper also describes the SIA papers and identifies a series of future research in SIA, KDD and Data Mining.

Keywords—Statistical Implicative Analysis; Knowledge Discovery Database; data mining, systematic review, mapping review.

I. INTRODUCTION

In recent years, our capacity to generate, transform, store, analyze and visualize data has increased, basically due to the high processing power and the low cost of the machines. However, within these huge and different types of data there is a lot of unknown information. The discovery of this information is possible thanks to Data Mining (DM), which among other sophisticated techniques applies artificial intelligence to find patterns and relationships within the data allowing the creation of models that are abstract representations of reality. Common tasks in Knowledge Discovery in Databases (KDD) are rule induction, classification and clustering problems, recognition pattern, predictive modeling, dependency detection, and so on [1].

A. The KDD process for extracting knowledge

The first KDD workshop was in August 20, 1989, Detroit MI, USA, enabling researchers and practitioners to gather around KDD. We define the KDD process as [2]: “The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”.

Data mining (DM) is a step in the general process to constituting the KDD process (see Figure 1). Data mining allows us to use specific algorithms for extracting patterns (including classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis) from data. Table I shows steps and subprocess in the KDD process [3]:

<table>
<thead>
<tr>
<th>Step</th>
<th>Subprocess</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>Learning the application domain</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>Creating a target dataset</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>Preprocessing</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>Data cleaning and preprocessing</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>Transformation</td>
</tr>
<tr>
<td>6</td>
<td>F</td>
<td>Data mining</td>
</tr>
<tr>
<td>7</td>
<td>G</td>
<td>Choosing the data mining algorithm(s)</td>
</tr>
<tr>
<td>8</td>
<td>H</td>
<td>Using discovered knowledge</td>
</tr>
</tbody>
</table>

B. Statistical Implicative Analysis and the Knowledge discovery

Statistical Implicative Analysis (SIA) was created by Regis Gras [4], thirty eight years ago and has a set of data analysis tools that that allows us to approach knowledge on the basis of the information contained in the database (individuals and variables). The approach is performed starting from the generation of asymmetric rules [5] between variables and variables classes, represented by tables (clusters non-hierarchical) [6], graphs (association rules) [7] and dendograms (hierarchical clusters, hierarchical oriented clusters) [8]. The statistical theory [9] and application of SIA are in continuous expansion and development. The SIA software tool is called CHIC [10] [11], the last Windows version is 7.0 and the CHIC free multiplatform version is called RCHIC [12]. SIA has an international group of active researchers since 2000 [13]. Usual CHIC functions are: Similarity Tree, Implicative Graph, Cohesion Tree and Reduction. Some of the complementary options implemented in CHIC are: the entropy is used when analyzing a large data sample; the supplementary variables are qualitative variables such as gender, education level or economic category; the contribution is used to know what are the subjects or classes of subjects more responsible for computed implications and the typicality indicates the typical subjects of the population for computed implications.

Rubén A. Pazmiño-Maji
Escuela Superior Politécnica de Chimbó
Riobamba, Ecuador
Email: rpazmino@espoch.edu.ec

Francisco J. García-Peña
Department of Computer Science
University of Salamanca
Salamanca, Spain
Email fgarcia@usal.es

Miguel A. Conde-González
Department of Computer Science
University of León
León, Spain
Email: miguel.conde@unileon.es

Statistical Implicative Analysis Approximation to KDD and Data Mining: A Systematic and Mapping Review in Knowledge Discovery Database Framework

The aim of this paper is to describe SIA papers in the last 5 years, to determine the position of SIA in KDD and DM, and identify a series of future research.

C. Paper framework

The approximation framework of Statistical Implicative Analysis to Knowledge Discovery and Data Mining are the KDD definition and the KDD steps proposed by Usama Fayyad [3]. This approximation allowed us to propose a new definition of SIA and to determine the steps that constitute the SIA process. We emphasize that this paper is a first approximation of SIA to Data Mining.

Section II describes the systematic and mapping review of literature and the steps in the research realized. Section III describes the results and their discussion. Finally, Section IV describes the conclusions.

II. METHOD

In this section we show in detail the steps of the methodology used.

A. Systematic and mapping review of literature

In the planning of systematic and mapping review the objectives were identified and the protocol was defined [14] [15]. The Protocol shows the method used in the systematic review and mapping to minimize the bias of researchers and that the methodology can be reproduced. Below we summarize the protocol used:

B. Research questions

The systematic mapping aims to answer these questions:

MQ1: What are the SIA papers by countries?
MQ2: Which are the SIA papers types?
MQ3: Which are the SIA paper Areas?
MQ4: What are the SIA papers tendency?

The systematic review aims to answer the question below:

RQ1: Which KDD step is the closest to SIA papers?
RQ2: Which KDD step is the furthest from SIA papers?
RQ3: How close are the SIA papers to KDD steps?
RQ4: How close are the SIA papers to KDD subprocess?
RQ5: How close are the SIA papers to Data Mining steps?
RQ6: Can you define the SIA based KDD definition?
RQ7: Can you define the process for SIA based KDD?

C. PICOC method

The paper of Petticrew and Roberts [16], proposed the PICOC method to define our scope:

- Population (P): Statistical implicative analysis papers in last five years (2012-2016).
- Intervention (I): SIA papers with explicit analysis process, in last five years (2012-2016).
- Comparison (C): No comparison intervention.
- Outcomes (O): SIA approximation percentages to KDD stages
- Context (C): SIA computational solutions.

D. Time period

The last 5 years (2012 to 2016)

E. Sources

The search was done in the following bibliographic databases [17]:

- EBSCO [18],
- Google Scholar [19],
- IEEEExplore [20],
- ProQuest [21],
- Scopus [22],
- WOS [23],
- Web of Science [24],
- ACM [25].

To answer the research questions raised, the inclusion and exclusion criteria were defined. They also allowed us to select the source SIA papers.
F. Inclusion and exclusion criteria

The inclusion criteria (IC) [26] are presented below:

IC1: The papers used a SIA methods real application
IC2: The proposed solution is applied on specialized software (Chic, Rchic, etc.)
IC3: The SIA process application is possible to make explicit
IC4: The papers are written in English language
IC5: The papers are reported in peer reviewed Workshop or Conference or Journal or Technical Reports

The exclusion criteria are presented below:

EC1: The paper is essentially theoretical, historical or a literature review
EC2: The SIA process application is not possible to make explicit
EC3: The SIA papers analysis methods do not use computer programs
EC4: The papers are written in Spanish, Italian, Portuguese or French language.

G. Search string

The group of primary studies were defined [27]. The final search string was described as follows: (“statistical implicative analysis” OR SIA) AND (LIMIT-TO (PUBYEAR, 2016) OR (LIMIT-TO (PUBYEAR, 2015) OR (LIMIT-TO (PUBYEAR, 2014) OR (LIMIT-TO (PUBYEAR, 2013) OR (LIMIT-TO (PUBYEAR, 2012)))) [28, 29] showed studies on control, if the search chain found relevant studies.

H. Review and mapping steps

The review and mapping were carried out and the details for this step are presented in the following subsection. Figure 2 shows the steps of systematic and mapping review with SIA papers.

![Figure 2. SIA papers mapping and review process](image)

According to Kitchenham [30] and literature mapping [31], quality checklists should be made. These checklists support the selection process. In this way, we produced the following checklist of quality.

I. Quality assessment

The quality assessment questions are presented below in Table II:

<table>
<thead>
<tr>
<th>Questions</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Are the SIA research goals clearly specified?</td>
<td>Yes=1 No=0 Half=0.5</td>
</tr>
<tr>
<td>2. Are the research aims achieved?</td>
<td>Yes=1 No=0 Half=0.5</td>
</tr>
<tr>
<td>3. Are the used data clearly described and their selection justified?</td>
<td>Yes=1 No=0 Half=0.5</td>
</tr>
<tr>
<td>4. Are the pre-processed data in SIA papers clearly described?</td>
<td>Yes=1 No=0 Half=0.5</td>
</tr>
<tr>
<td>5. Are the transformed data in SIA papers clearly described?</td>
<td>Yes=1 No=0 Half=0.5</td>
</tr>
<tr>
<td>6. Are the SIA's papers algorithms clearly described and their selection justified?</td>
<td>Yes=1 No=0 Half=0.5</td>
</tr>
<tr>
<td>7. Are the SIA's papers methods clearly described and their selection justified?</td>
<td>Yes=1 No=0 Half=0.5</td>
</tr>
<tr>
<td>8. Is the data analysis process done by the computer?</td>
<td>Yes=1 No=0 Half=0.5</td>
</tr>
<tr>
<td>9. How clear are the links between data, transformed data, analysis, interpretation and conclusions?</td>
<td>Yes=1 No=0 Half=0.5</td>
</tr>
</tbody>
</table>

III. Result and Discussion

In this section we show the results obtained in the mapping and systematic review process.

A. Mapping literature review

In this section we describe the results and their discussion about mapping and systematic literature review.

1) What are the SIA papers by countries?

![Figure 3. Percentage of SIA papers by countries](image)

Figure 3 shows that the most frequent countries in the selected literature are from France (40.0%) and Poland (16.7%). France is the most frequent country because SIA
theory was originated with Regis Gras born in France.

2) Which are the SIA papers types?

Figure 4 illustrates that studies from the chosen literature are 73% Articles, 20% Conference Papers, and 7% Book Chapter Books. This is because SIA international congress is producing new papers every two years.

3) Which are the SIA papers areas?

Figure 5 illustrates that most of the studies have targeted

Computer Science (32.0%), Social Sciences (24.0%), Mathematics (23.6%) and Decision Sciences (6%). The four areas added are approximately 80%, this is because SIA theory was originated on didactic, mathematics and his methods are used in computer science.

4) What are the SIA papers tendency?

Figure 6 shows the tendency of SIA papers in general (in red) and SIA real applications papers (in blue), in the last five years. SIA real applications papers tend to increase because there are fewer the theoretical, historical or a literature review papers than real applications papers.

B. Systematic literature review

In this section we describe the results and their discussion about systematic review.

1) Which KDD step is the closest to SIA papers?

Steps 1 (Learning the application domain) and steps 2 (Creating a target dataset) with 100 % both, are the steps closest to KDD steps. This is because, all SIA real applications papers need a goal and a target dataset.

2) Which KDD step is the furthest from SIA papers?

Step 7 (Data mining) with 31 %, is the step furthest to KDD steps. This is because, most papers showed not a Data Mining pattern clearly, had descriptive analysis, histograms, bar diagrams, box plots and SIA graphics like a similarity tree, cohesion tree, implicative graph or reduction.

3) How close are the SIA papers to KDD steps?

Table III shows the reference of 35 SIA papers, the compliance or not of KDD steps process 1 to 9 (see introduction, quality assessment and [18]). Also, Table III shows in the last two columns the numbers of positive compliance and the respective percentage. The final percentage is the 74,6 %, which is medium high close to KDD steps process. This means that KDD steps process and SIA steps process are the same in a 74,6 %.
Table III: SIA Papers & KDD Steps Related to Table I

<table>
<thead>
<tr>
<th>SIA papers</th>
<th>KDD steps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 N %</td>
</tr>
<tr>
<td>[32]</td>
<td>1 1 0 0 1 1 1 0 1 1 6 66.7</td>
</tr>
<tr>
<td>[33]</td>
<td>1 1 1 1 1 1 1 1 1 1 8 88.9</td>
</tr>
<tr>
<td>[34]</td>
<td>1 1 0 0 1 1 1 1 1 1 7 77.8</td>
</tr>
<tr>
<td>[35]</td>
<td>1 1 0 0 1 0 0 0 0 0 3 33.3</td>
</tr>
<tr>
<td>[36]</td>
<td>1 1 0 0 1 1 1 1 1 1 6 66.7</td>
</tr>
<tr>
<td>[37]</td>
<td>1 1 0 0 1 0 0 0 1 1 5 55.6</td>
</tr>
<tr>
<td>[38]</td>
<td>1 1 1 1 1 1 1 1 1 1 9 100.0</td>
</tr>
<tr>
<td>[39]</td>
<td>1 1 1 0 1 0 0 1 0 0 5 55.6</td>
</tr>
<tr>
<td>[40]</td>
<td>1 1 0 1 1 1 1 1 1 1 8 88.9</td>
</tr>
<tr>
<td>[41]</td>
<td>1 1 1 1 1 1 1 0 1 1 6 66.7</td>
</tr>
<tr>
<td>[42]</td>
<td>1 1 1 1 1 1 1 0 1 1 6 66.7</td>
</tr>
<tr>
<td>[43]</td>
<td>1 1 1 1 1 1 1 0 1 1 7 77.8</td>
</tr>
<tr>
<td>[44]</td>
<td>1 1 1 1 1 1 1 0 1 1 8 88.9</td>
</tr>
<tr>
<td>[45]</td>
<td>1 1 1 0 1 1 0 1 1 1 8 88.9</td>
</tr>
<tr>
<td>[46]</td>
<td>1 1 1 1 1 1 1 1 1 9 100.0</td>
</tr>
<tr>
<td>[47]</td>
<td>1 1 0 0 1 1 0 1 1 6 66.7</td>
</tr>
<tr>
<td>[48]</td>
<td>1 1 1 1 1 1 1 1 1 4 44.4</td>
</tr>
<tr>
<td>[49]</td>
<td>1 1 0 1 1 0 0 0 0 5 55.6</td>
</tr>
<tr>
<td>[50]</td>
<td>1 1 1 1 1 1 0 1 1 8 88.9</td>
</tr>
<tr>
<td>[51]</td>
<td>1 1 0 1 1 1 0 0 0 5 55.6</td>
</tr>
<tr>
<td>[52]</td>
<td>1 1 1 1 1 1 0 1 1 8 88.9</td>
</tr>
<tr>
<td>[53]</td>
<td>1 1 1 0 1 0 0 1 1 6 66.7</td>
</tr>
<tr>
<td>[54]</td>
<td>1 1 1 1 0 0 1 1 1 6 66.7</td>
</tr>
<tr>
<td>[55]</td>
<td>1 1 0 1 1 1 0 1 1 6 66.7</td>
</tr>
<tr>
<td>[56]</td>
<td>1 1 1 1 1 1 0 1 1 8 88.9</td>
</tr>
<tr>
<td>[57]</td>
<td>1 1 1 1 1 0 0 1 1 7 77.8</td>
</tr>
<tr>
<td>[58]</td>
<td>1 1 1 0 1 0 0 1 1 6 66.7</td>
</tr>
<tr>
<td>[59]</td>
<td>1 1 1 1 1 1 1 1 1 9 100.0</td>
</tr>
<tr>
<td>[60]</td>
<td>1 1 1 1 1 1 0 1 1 8 88.9</td>
</tr>
<tr>
<td>[61]</td>
<td>1 1 1 1 1 1 1 1 1 9 100.0</td>
</tr>
<tr>
<td>[62]</td>
<td>1 1 1 1 1 1 0 0 0 4 44.4</td>
</tr>
<tr>
<td>[63]</td>
<td>1 1 1 1 1 1 0 0 1 1 7 77.8</td>
</tr>
<tr>
<td>[64]</td>
<td>1 1 0 1 1 1 0 1 1 6 66.7</td>
</tr>
<tr>
<td>[65]</td>
<td>1 1 1 0 1 0 0 1 1 6 66.7</td>
</tr>
<tr>
<td>[66]</td>
<td>1 1 1 0 1 0 0 1 1 6 66.7</td>
</tr>
</tbody>
</table>

4) **How close are the SIA papers to KDD subprocess?**

Table III shows the reference of 35 SIA papers, the compliance or not of KDD subprocess A to E (see introduction, quality assessment and [18]). Table IV shows in the last two columns the numbers of positive compliance and the respective percentage. The final percentage 94.2%, means that KDD and SIA subprocess are very similar.

5) **How close are the SIA papers to the Data Mining steps?**

Observing step 5 (Choosing the function of data mining, 94%), step 6 (Choosing the data mining algorithm, 54%), step 7 (Data mining, 31%) and subprocess (Data mining, 86%) we have 66.4% of approach to Data Mining process. It is a medium high percentage. This is because the SIA methods can be similar to the Data Mining methods.

6) **Can you define the SIA process based on KDD process?**

The SIA definition based on the KDD process could be:

Statistical Implicative Analysis is the process of identifying valid, useful, and understandable, r-rules in data.

In the previous SIA definition, the words valid, useful, and understandable depend on the researcher interpretation of the SIA methods used and the results obtained. The novel word is not in a SIA definition because the use of SIA methods is novel in Data Mining.

7) **Can you define the process for SIA based KDD?**

Figure 7 shows the steps graphic constituting the SIA process (based in KDD).
IV. CONCLUSIONS

The aim of this paper is to describe SIA papers in last 5 years (2012-2016), to determine the approximation of SIA to KDD and DM, and identify a series of future research. To describe SIA papers, we use the mapping and systematic literature review methods. The most frequent country in the selected literature is France (40%); also 73% of the chosen literature are Articles. The studies were targeted in Computer Science (32.0%), Social Sciences (24.0%), Mathematics (23.6%), and Decision Sciences (6%). The four areas are approximately 80% of SIA papers total areas.

It is important to note that the results obtained depend closely on the approximation framework used, in this case the steps of the Faday process. It was determined that SIA and KDD are strongly related, with a contention relationship of SIA to KDD. In the first approximation, considering all steps without subprocesses, a contention of 74.6% corresponding to medium high approximation was obtained. In the second approach, considering all the subprocesses a contention of SIA to KDD of 94.2% was obtained that is high approximation to KDD process. This is summarized by indicating that the overall approximation rate considering the complete Faday process is 84.4% which is medium high approximation. The approximation achieved allowed to propose the steps that constitute the SIA process analysis, besides a new definition. The process D and steps 5 and 6 allowed us to give a first approximation of the SIA to data mining, obtaining 66.4% of contention between SIA and Data Mining corresponding to medium high percentage.

The future research in SIA, KDD and Data Mining can be to answer the following questions: Do SIA methods work with big data? Can we use similarity tree and cohesion tree like a data mining hierarchical cluster method? Can we use implicative graph like a data mining rule induction method? And Can we use reduction like a data mining cluster method?. For example we have the questions: Do SIA methods work with big data? Can we use similarity tree and cohesion tree like a data mining hierarchical cluster method? Can we use reduction SIA method like a data mining cluster method? And Can we use implicative graph like a data mining rule induction method?. The answer to the abode questions is future research in SIA, KDD and Data Mining.

ACKNOWLEDGMENT

We would like to thank the University of Salamanca PhD programme on Education in the Knowledge Society scope. Similarly, we want to thank Escuela Superior Politécnica de Chimborazo for funding to perform this research.

REFERENCES


