Advances in Intelligent Systems and Computing 407

Tarek Gaber
Aboul Ella Hassanien
Nashwa El-Bendary
Nilanjan Dey *Editors*

The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28–30, 2015, Beni Suef, Egypt

Springer

Advances in Intelligent Systems and Computing

Volume 407

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

More information about this series at http://www.springer.com/series/11156

Tarek Gaber · Aboul Ella Hassanien
Nashwa El-Bendary · Nilanjan Dey
Editors

The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28–30, 2015, Beni Suef, Egypt

Springer

Editors
Tarek Gaber
Faculty of Computers and Informatics
Suez Canal University
Ismailia
Egypt

Aboul Ella Hassanien
Faculty of Computers and Information
Cairo University
Giza
Egypt

Nashwa El-Bendary
Arab Academy for Science
Giza
Egypt

Nilanjan Dey
Bengal College of Engineering and Technology
Durgapur, West Bengal
India

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-26688-6 ISBN 978-3-319-26690-9 (eBook)
DOI 10.1007/978-3-319-26690-9

Library of Congress Control Number: 2015955384

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This edited volume of *Advances in Intelligent Systems and Computing* contains the accepted papers presented at the main track of the 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28–30, 2015, Beni Suef, Egypt.

The aim of AISI2015 is to provide an internationally respected forum for academics, researchers, analysts, industry consultants, and practitioners in the fields involved to discuss and exchange recent progress in the area of informatics and intelligent systems technologies and applications. The conference has three major tracks, namely Intelligent Systems, Intelligent Robotics Systems, and Informatics.

The plenary lectures, tutorials, and the progress and special reports bridged the gap between the different fields of machine learning and informatics, making it possible for nonexperts in a given area to gain insight into new areas. Also included among the speakers were several young scientists, namely postdocs and students, who brought new perspectives to their fields. We expect that the future AISI conferences will be as stimulating as this most recent one was, as indicated by the contributions presented in this proceedings volume.

The first edition of AISI2015 was organized by the Scientific Research Group in Egypt (SRGE) jointly by Faculty of Computers and Information, Beni Suef University, Beni Suef, Egypt. The conference was organized under the patronage of Prof. Amin Lotfy, President of Beni Suef University. The conference received 85 papers from 29 countries and accepted 43 papers.

The organization of the AISI2015 conference was entirely voluntary. The reviewing process of AISI2015 required huge effort from the International Technical Program Committee. Each paper has been reviewed by at least three reviewers. Therefore, we would like to thank all the members of this committee for their contribution to the success of the AISI2015 conference. Also, we would like to express our sincere gratitude and appreciation to the host of AISI2015, Beni Suef

University, Beni Suef, Egypt, and to the publisher, Springer, for their hard work and support in the organization of the conference. Last but not least, we would like to thank all the authors of this proceedings for their high-quality contributions.

The friendly and welcoming attitude of the AISI2015 conference supporters and contributors made this event a success!

Egypt Tarek Gaber
Egypt Aboul Ella Hassanien
Egypt Nashwa El-Bendary
India Nilanjan Dey
September 2015

Contents

Part I
Intelligent Systems and Informatics (I)

Automatic Rules Generation Approach for Data Cleaning in Medical Applications

Asmaa S. Abdo, Rashed K. Salem and Hatem M. Abdul-Kader

Abstract Data quality is considered crucial challenge in emerging big data scenarios. Data mining techniques can be reutilized efficiently in data cleaning process. Recent studies have shown that databases are often suffered from inconsistent data issues, which ought to be resolved in the cleaning process. In this paper, we introduce an automated approach for dependably generating rules from databases themselves, in order to detect data inconsistency problems from large databases. The proposed approach employs confidence and lift measures with integrity constraints, in order to guarantee that generated rules are minimal, non-redundant and precise. The proposed approach is validated against several datasets from healthcare domain. We experimentally demonstrate that our approach outperform significant enhancement over existing approaches.

Keywords Data quality · Data mining · Data inconsistency · Data cleaning · Electronic medical records (EMR)

1 Introduction

Data is most important value in today's economy [1]. The value of data highly depends on its degree of quality. Henceforward, existence of inconsistency issues in data intensely decreases their assessment, making them misinformed, or even harmful.

A.S. Abdo (✉) · R.K. Salem · H.M. Abdul-Kader
Information System Department, Faculty of Computers and Information Menoufia University, Menoufia, Egypt
e-mail: asm2009_asm2009@yahoo.com

R.K. Salem
e-mail: rdkhalil@yahoo.com

T. Gaber et al. (eds.), *The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28–30, 2015, Beni Suef, Egypt*,
Advances in Intelligent Systems and Computing 407, DOI 10.1007/978-3-319-26690-9_1

The quality of data is an increasingly pervasive problem, as data in real world databases quickly degenerates over time and effects the results of the mining. Such poor data quality often emerges due to violations of integrity constraints, which results in incorrect statistics, and ultimately wasting of time and money [2, 3].

Companies lose billions of dollars annually due to poor data quality [4]. As a result, detecting inconsistent data is very important task in the data cleaning process. Doubtless, ensuring high quality dependable data is a competitive advantage to all businesses, which requires accurate data cleaning solutions [5, 6].

Data cleaning refers to the process of maintaining corrupted and/or inaccurate records. This process is mandatory in data management cycle before mining and analyzing data [1].

However, most of the existing data cleaning techniques in the literature focus on record matching [4], which match master cleaned records with a probably imprecise records. It is still necessary to tackle the problem of data inconsistency with the help of data themselves and without the need for external master copy of data. During resolving data inconsistencies, several integrity constraints are ensured, e.g., Functional Dependencies (FDs) and Conditional Functional Dependencies (CFDs) [7, 8].

Medical application domain is one of the most critical applications that suffers from inconsistent and dirty data issues. As ensuring data quality of electronic medical records is very important in healthcare data management purposes, whereas critical decisions are based on patient status apprised from medical records [9–11]. Herein, we are interested to generate data quality rules, which then used for resolving data inconsistencies in such medical databases.

Indeed, the main contribution of this paper is to propose an approach for generating dependable data cleaning rules. Such discovered rules are exploited not only for detecting inconsistent data, but also for correcting them. The proposed approach utilizes data mining techniques for discovering dependable rules, it bases mainly on frequent closed-patterns and their associated generators to their closure that speed up rules generation process. The experimental results conducted over medical datasets verify the effectiveness and accuracy of the proposed approach against CCFD-ZartMNR algorithm [12].

This paper is organized as follows: Sect. 2 discusses related works. In Sect. 3 present data quality in medical applications. Section 4 introduces the proposed approach for generating dependable data cleaning rules. Section 5 discusses the experimental study and results conducted for different medical datasets. Finally, Sect. 6 concludes the proposed work and highlight future trends.

2 Related Work

Despite the urgent need for precise and dependable techniques for enhancing data quality and data cleaning problems, there is not vital solution up to now to these problems. There has been little discussion and analysis about enhancing data inconsistency. However, most of recent work focus on record matching and duplicate detection [13].

Database and data quality researchers have discussed variety of integrity constraints based on Functional Dependencies (FD) [7, 14–16]. These constraints are developed essentially for schema design, but are often not able to embed the semantic of data for data cleaning.

Other researchers focuses on extension of FD, they have proposed what is so-called Conditional Functional Dependencies (CFD) and Conditional Inclusion Dependencies (CID). Such integrity constraints are employed for capturing errors in data [8, 17, 18].

Moreover, several data quality techniques are proposed to clean missy tuples in databases [19]. Statistical inference approaches are studied in [20]. These approaches tackle missing values in order to enhance quality of data. From technological part, there are several open source tools which developed for handling messy data [21].

Besides, data transformation methods such as commercial ETL (Extract, Transformation and Loading) tools are developed for data cleaning [22].

The usage of editing rules in combination with master data is discussed in [23]. Such rules are able to find certain fixes by updating input tuples with master data. However, this approach requires defining editing rules manually for both relations, i.e., master relation and input relation, which is very expensive and time-consuming.

Furthermore, lots of work are proposed in the literature relying on domain specific similarity and matching operators, such works include record matching, record linkage, duplicate detection, and merge purge, [13, 24, 25]. These approaches define two functions; namely match and merge [26]. While match function identifies duplication of records, the merge function combines the two duplicated records into one.

From the literature, we reutilized CFD as a special case of association rules [8, 12]. The relationship between minimal constant CFD and item set mining, association rule as similar relationship to CFD that work on minimal non redundant rules [27].

3 Data Quality in Medical Applications

In era of Electronic Medical Records (EMR), which are objects of knowledge about patients medical and clinical data [28]. Healthcare stakeholders and providers of health care services need such information and knowledge not only at the point of service but also at the point of clinical treatment decisions for improving health care quality [29, 30]. They require such knowledge with precise quality to maximize the benefit of decision-making process [31]. However, maintaining exact and reliable information about diseases treated is based on precise data stored about patient [32]. Therefore, clinical and health service research aimed at accurate, reliable and

Table 1 Records from thyroid (hypothyroid) dataset

Patient number	Pregnant	Hypopituitary	On antithyroid medication	TBG measured
98	f	f	t	f
162	f	f	f	f
164	f	f	f	f
175	f	f	t	f
183	f	f	f	f
214	f	f	t	f
218	f	f	f	f
261	f	f	f	f
742	f	f	t	f
1253	f	f	f	f

complete statistical information about the uses of health care services within a community.

Example 1 Consider a sample from thyroid (hypothyroid) dataset. The following relation schema *R* is about patient data given in Table 1. Technicians know that these data suffer from inconsistency. The proposed approach aims to discover dependable data quality rules for detecting such inconsistencies.

4 Proposed Approach

The main purpose of the proposed approach is to discover interest minimal non re-dundant constant CFD rules that cover all set of rules. In other words, the discovered rules are minimal and complete with respect to specified support and confidence thresholds. Such discovered rules are employed for dirty data identification and treating data inconsistencies. The proposed approach relies on generating closed frequent patterns and their associated generators according. The associated generators are defined as the closure of closed frequent patterns from all set of frequent patterns.

Given an instance r of a relation schema *R*, support *s*, and confidence *c* thresholds, the proposed approach discovers Interest-based Constant Conditional Functional Dependencies (ICCFDs), abbreviated ICCFD-Miner approach. The discovered ICCFDs ensure finding interest minimal non-redundant dependable Constant Conditional Functional Dependencies (CCFDs) rules with constant patterns in *r*.

The flow of two main steps for generating minimal non redundant rules is shown in Fig. 1. The proposed approach is detailed as follows:

Input: Dataset and two predefined thresholds, i.e., minimum support (minsup), minimum confidence (minconf), are the input to the ICCFD-Miner approach.

Fig. 1 The proposed ICCFD-Miner approach

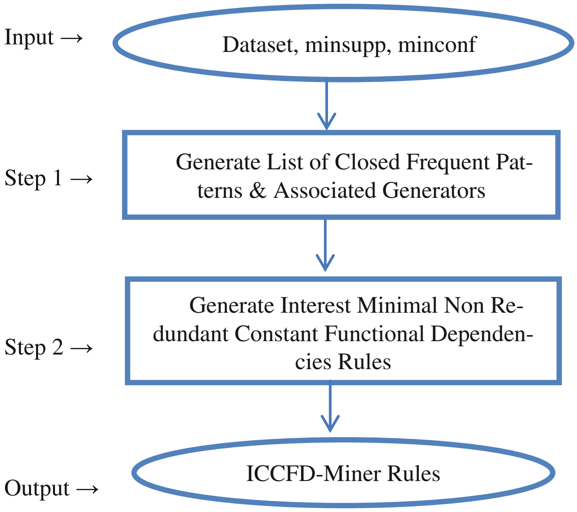
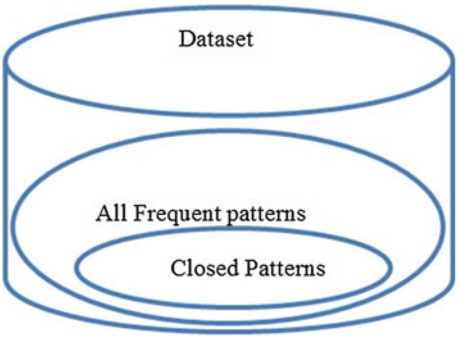


Fig. 2 Search space domain



Step 1: Given user defined minimum support threshold, the list of closed frequent patterns and their associated generators to their closure can be generated. In order to minimize the search space and saving the time of rules generation, The proposed approach utilize such generated closed frequent patterns and their associated generators instead of working on generating all frequent patterns [12]. The search space domain is indicated in Fig. 2.

Let us define Support CFD φ: (X → A), as the number of records in the dataset that contain XUA to the total number of records in the database. Support threshold based on the idea that values that occur together frequently have more evidence to validate that they are correlated. Support of a CFD φ: (X → A) where *X* generator pattern and *A* is (closed/generator) [33], defined as

$$\text{support}(X \rightarrow A) = \frac{\text{Number of tuples containning values in X and A}}{\text{Total number of tuples in relation}} \quad (1)$$

Step 2: Given user defined confidence threshold. The set of interest minimum non redundant constant conditional functional dependencies data quality rules are generated. While the literature utilizes only support and confidence for generating such rules, the proposed approach consider interest measure into account for generating more dependable and reliable rules. The form of rules for each frequent generator pattern *X* finds its proper supersets *A* from set of frequent closed patterns. Then, from *X* and *A* add rule antecedent (Generator) → consequence (closed/generator) as φ: X → A.

Let us define Confidence CFD as the number of records in the dataset that satisfy CFD divided by number of records that satisfy left hand side of rule.

$$\text{confidence}(X \rightarrow A) = \frac{\text{support}(\phi)}{\text{support}(x)} \quad (2)$$

Confidence measures reliability of rule, since the value of confidence is real number between 0 and 1.0 [34]. The pitfall of Confidence is that ignores support of right hand side of rules. As consequence, we add data quality measure called Interest (Lift) which generates more dependent rules when defined it as greater than one.

Let us define *Lift CFD* as measuring the degree of compatibility of left hand side and right hand side of rules as, i.e., occurrence of both left hand side and right hand side [35]. We set here Lift value > 1 to obtain dependent ICCFD-Miner rules. For example lift of this CFD rule ϕ: (X → A) is

$$\text{lift}(X \rightarrow A) = \frac{\text{confidence}(\phi)}{\text{support}(A)} \quad (3)$$

This approach optimizes the process of the rules generation compared with the most related methods. Let us addressing an example of generated rule form from dataset utilized in Sect. 3 as follows:

ϕ: pregnant = f, hypopituitary = f → on antithyroid medication = f, TBG measured = f

[support = 0.98 (3679/3772) confidence = 0.99 lift > 1].

"If she is not pregnant and not has hypopituitary then it must not take any antithyroid medication and its TBG measured value should be false".

This consider very important rule to physicians when determining drug to patient. Moreover, such rules are used as critical review to decision maker which determine total number of patient that take/not take antithyroid medication. Tuples that violate this generated rule are shown in Table 2. We highlight detected error values in data.

Table 2 Records that violate generating rule

Patient number	Pregnant	Hypopituitary	On antithyroid medication	TBG measured
98	f	f	t	f
162	f	f	f	f
164	f	f	f	f
175	f	f	t	f
183	f	f	f	f
214	f	f	t	f
218	f	f	f	f
261	f	f	f	f
742	f	f	t	f
1253	f	f	f	f

5 Experimental Study

Using two real-life datasets, we evaluate our proposed approach described in Sect. 4. In order to assess their performance in current domain of real life application especially in critical application such as medical applications. The exploited datasets have large amount of information about patients and their status. The proposed approach used for generating dependable rules in these datasets in order to enhance their data quality. The cleaned data become on demand data access for decision maker in healthcare systems, enabling accurate decisions based on their exact quality data.

We evaluate the following factors on the efficiency and the accuracy of ICCFD-Miner rules produced such threshold support (*sup*), confidence (*conf*), size of sample relation r (the number of instances in r), arity of relation r (the number of columns in r), and time complexity.

5.1 Experimental Setting

The experiments are conducted using two real-life datasets about diseases taken from the UCI machine learning repository (http://archive.ics.uci.edu/ml/) namely, Thyroid (hypothyroid), primary-tumor. Table 3 shows the number of attributes and the number of instances for each dataset.

The proposed approach is implemented using java (JDK1.7). The implementation is tested on machine equipped with Intel(R) Pentium(R) Dual CPUT3400 @ 2.16 GHz 2.17 GHz processor with 2.00 GB of memory running on windows 7 operating system. The proposed approach runs mainly in main memory. Each experiment is repeated at least five times and the average reported here.

Table 3 Datasets description

Dataset name	Arity (Number of columns)	Size (Number of instances)
Thyroid(hypothyroid)	30	3772
primary-tumor	18	339

5.2 Experimental Results

Now, we show and discuss the results on real world dataset described in previous section. Note that we aim to evaluate the effectiveness of rules generation of the proposed approach against CCFD-ZartMNR algorithm [12]. Experiments show that proposed approach always produce less number of rules but more accurate. The generated rules are interest-based minimal and non-redundant. Experiments also show that ICCFD-Miner outperforms the other algorithm with respect to time for rule generation.

Experiment-1: In this experiment, rules are generated from thyroid (hypothyroid) dataset. This data set contains 30 attributes, 3772 records of patient data describing patient information about the hypothyroid diagnoses data.

By varying values of support (sup) and confidence (conf) thresholds as shown in Fig. 3, we notice that the proposed approach always generate accurate interest minimum non redundant rules compared to CCFD-ZartMNR algorithm. For example in Fig. 3, at minimum support = 0.97 and minimum confidence = 0.99 the number of rules generated from the proposed ICCFD-Miner approach = 85 rules compared to output generated rules from CCFD_ZartMNR = 220 rules. Results from Fig. 4 shows that the proposed approach generates rules at different sup, conf values in less time compared to CCFD-ZartMNR algorithm. For example in Fig. 4, at minimum support = 0.97 and minimum confidence = 0.99 response time of the proposed ICCFD-Miner approach = 312 ms compared to response time of existing approach CCFD_ZartMNR = 368 ms.

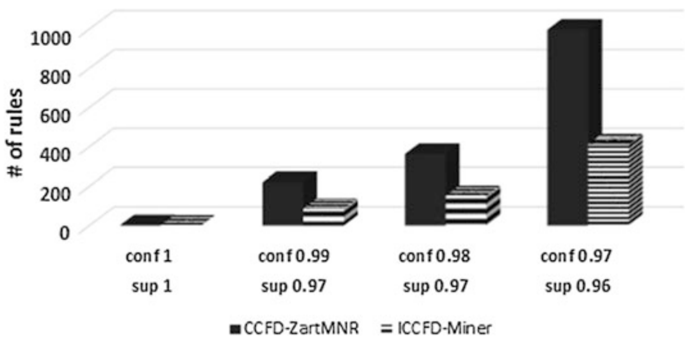


Fig. 3 Total number of rules generated of thyroid (hypothyroid) dataset

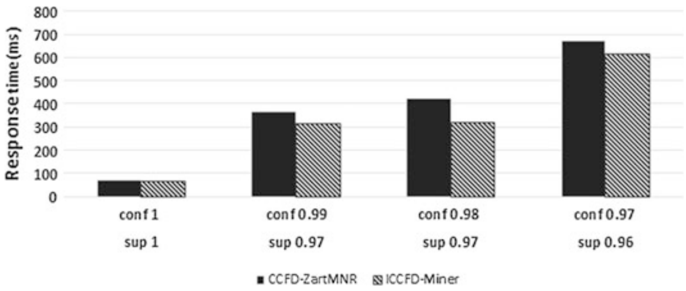


Fig. 4 Response time measure about thyroid (hypothyroid) dataset

Experiment-2: The experimental results conducted over primary-tumor disease dataset are shown in Figs. 5 and 6. This data set contains 18 attributes, 339 records describing patient information about the on primary-tumor disease diagnoses data.

Figures 5 and 6 validate the efficiency of the proposed approach against CCFD-ZartMNR algorithm in both number of rules generated and response time measure.

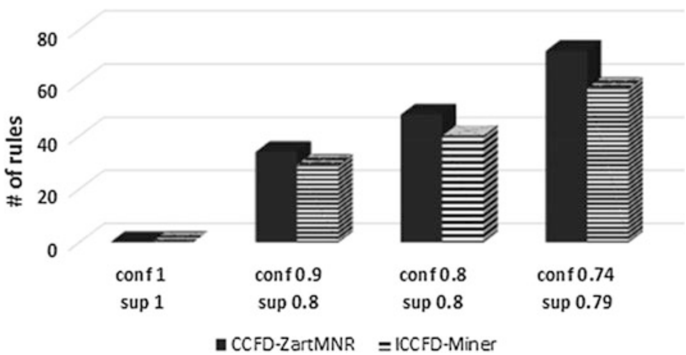


Fig. 5 Total number of rules generated of primary-tumor dataset

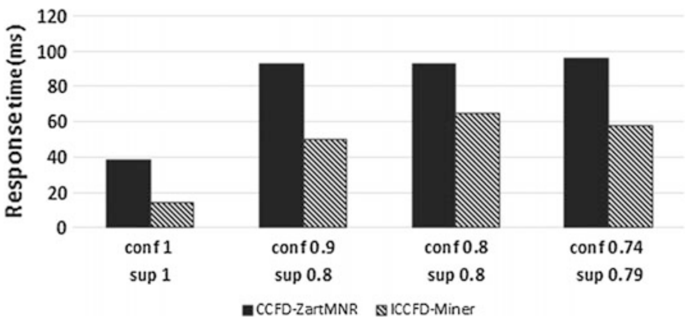


Fig. 6 Response time measure about primary-tumor dataset

Finally, we believe that the proposed approach, i.e., ICCFD-Miner, outperforms CCFD_ZartMNR due to its implying lift measure when generating dependable rules. Furthermore, the proposed approach focus on closed patterns with their associated generators, i.e., supersets of closed patterns, as a search space for generating the more accurate and reliable rules.

6 Conclusions

In this paper, we have presented ICCFD-Miner approach that discovers precise data quality rules for resolving data inconsistency problem. The proposed approach yields a promising method for detecting semantic data inconsistency. The main target of the proposed approach is keeping the database in consistent state. Generated rules are exploited as data cleaning solution to resolve inconsistency problem in current application domains. ICCFD-Miner relies on lift measure in addition to support and confidence measures for generating dependable minimal and non-redundant rules. The ICCFD-Miner is validated and evaluated over two real life datasets from medical domain. The experimental results confirm the effectiveness and usefulness of the proposed approach against CCFD_ZartMNR algorithm. The proposed approach perform well across several dimensions such as effectiveness, accuracy of number of rules generated, and run time. Finally, we plan to investigate a technique for fixing errors autonomously with generated rules from ICCFD-Miner.

References

1. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F.: Automatic synthesis of data cleansing activities (2011)
2. Li, J., Liu, J., Toivonen, H., Yong, J.: Effective pruning for the discovery of conditional functional dependencies. Comput. J. **56**, 378–392 (2013)
3. Yakout, M., Elmagarmid, A.K., Neville, J.: Ranking for data repairs. In: Proceeding—International Conference Data Engineering, pp. 23–28 (2010)
4. Fan, W., Li, J., Ma, S., Tang, N., Yu, W.: Interaction between record matching and data repairing. In: Proceeding Journal of Data and Information Quality (JDIQ) vol. 4(4), p. 16 (2014)
5. Wang, J., Tang, N.: Towards dependable data repairing with fixing rules. In: SIGMOD Conference, pp. 457–468 (2014)
6. Fan, W., Geerts, F.: Foundations of data quality management. Synth. Lect. Data Manage. **4**, 1–217 (2012)
7. Liu, J., Li, J., Liu, C., Chen, Y.: Discover dependencies from data—a review. IEEE Trans. Knowl. Data Eng. **24**, 251–264 (2012)
8. Vo, L.T.H., Cao, J., Rahayu, W.: Discovering conditional functional dependencies. Conf. Res. Pract. Inf. Technol. Ser. **115**, 143–152 (2011)
9. Rodríguez, C.C.G., Riveill, M., Antipolis, S.: e-Health monitoring applications : what about data quality ? (2010)

10. Mans, R.S., van der A., Wil M.P., Vanwersch, R.J.: Data Quality Issues. Process Mining in Healthcare, pp. 79–88. Springer, Berlin (2015)
11. Kazley, A.S., Diana, M.L., Ford, E.W., Menachemi, N.: Is electronic health record use associated with patient satisfaction in hospitals? Health Care Manage. Rev. **37**, 23–30 (2012)
12. Kalyani, D.D.: Mining constant conditional functional dependencies for improving data quality. **74**, 12–20 (2013)
13. Bharambe, D., Jain, S., Jain, A.: A survey : detection of duplicate record. **2**, (2012)
14. Cong, G., Fan, W., Geerts, F., Jia, X., Ma, S.: Improving data quality: consistency and accuracy. In: Proceeding 33rd International Conference Very Large Data Bases, pp. 315–326. Vienna, Au (2007)
15. Hartmann, S., Kirchberg, M., Link, S.: Design by example for SQL table definitions with functional dependencies. VLDB J. **21**, 121–144 (2012)
16. Yao, H., Hamilton, H.J.: Mining functional dependencies from data. Data Min. Knowl. Discov. **16**, 197–219 (2008)
17. Bohannon, P., Fan, W., Geerts, F., Jia, X., Kementsietsidis, A.: Conditional functional dependencies for data cleaning. In: Proceeding—International Conference Data Engineering, pp. 746–755 (2007)
18. Bauckmann, J., Abedjan, Z., Leser, U., Müller, H., Naumann, F.: Discovering conditional inclusion dependencies. In: 21st ACM International Conference on Information and Knowledge Management, pp. 2094–2098. (2012)
19. Fan, W., Geerts, F.: Capturing missing tuples and missing values. In: Proceeding 29th ACM SIGACT-SIGMOD-SIGART Symposium Principle of Database System, pp. 169–178 (2010)
20. Mayfield, C., Neville, J., Prabhakar, S.: ERACER: a database approach for statistical inference and data cleaning. In: Proceeding ACM SIGMOD International Conference Management Data, pp. 75–86 (2010)
21. Larsson, P.: Evaluation of open source data cleaning tools : open refine and data wrangler. (2013)
22. Vassiliadis, P., Simitsis, A.: Extraction, transformation, and loading. Encycl. Database Syst. 1095–1101 (2009)
23. Fan, W., Li, J., Ma, S., Tang, N., Yu, W.: Towards certain fixes with editing rules and master data. VLDB J. **21**, 213–238 (2012)
24. Fan, W., Gao, H., Jia, X., Li, J., Ma, S.: Dynamic constraints for record matching. VLDB J. **20**, 495–520 (2011)
25. Reiter, J.: Data quality and record linkage techniques. J. Am. Stat. Assoc. **103**(482), 881 (2008)
26. Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S.E., Widom, J.: Swoosh: a generic approach to entity resolution. VLDB J. **18**, 255–276 (2009)
27. Zaki, M.J.: Mining non-redundant association rules. Data Min. Knowl. Discov. **9**, 223–248 (2004)
28. Chang, I.-C., Li, Y.-C., Wu, T.-Y., Yen, D.C.: Electronic medical record quality and its impact on user satisfaction—Healthcare providers' point of view. Gov. Inf. Q. **29**, 235–242 (2012)
29. Weiskopf, N.G., Weng, C.: Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J. Am. Med. Inform. Assoc. 144–151 (2012)
30. Groves, P., Kayyali, B., Knott, D., Van Kuiken, S.: The " Big Data " Revolution in Healthcare. McKinsey, New York (2013)
31. Kush, R.D., Ph.D., Helton, E., Rockhold, F.W., Hardison, C.D.: Electronic health records, medical research, and the tower of Babel. 16–18 (2008)
32. Koh, H.C., Tan, G.: Data mining applications in healthcare. J. Healthc. Inf. Manage. **19**, 64–72 (2005)
33. Chiang, F., Miller, R.J.: Discovering data quality rules. In: Proceeding VLDB Endowment, pp. 1166–1177 (2008)

34. Medina, R., Nourine, L.: A unified hierarchy for functional dependencies, conditional functional dependencies and association rules. In: LNAI, Lecture Notes Computer Science (including Subseries Lecture Notes Artifical Intelligent Lecture Notes Bioinformatics). vol. 5548, pp. 98–113 (2009)
35. Hussein, N., Alashqur, A., Sowan, B.: Using the interestingness measure lift to generate association rules. J. Adv. Comput. Sci. Technol. **4**, 156 (2015)

Action Classification Using Weighted Directional Wavelet LBP Histograms

Maryam N. Al-Berry, Mohammed A.-M. Salem, Hala M. Ebeid, Ashraf S. Hussein and Mohamed F. Tolba

Abstract The wavelet transform is one of the widely used transforms that proved to be very powerful in many applications, as it has strong localization ability in both frequency and space. In this paper, the 3D Stationary Wavelet Transform (SWT) is combined with a Local Binary Pattern (LBP) histogram to represent and describe the human actions in video sequences. A global representation is obtained and described using Hu invariant moments and a weighted LBP histogram is presented to describe the local structures in the wavelet representation. The directional and multi-scale information encoded in the wavelet coefficients is utilized to obtain a robust description that combine global and local descriptions in a unified feature vector. This unified vector is used to train a standard classifier. The performance of the proposed descriptor is verified using the KTH dataset and achieved high accuracy compared to existing state-of-the-art methods.

Keywords 3D stationary wavelet · Action recognition · Local binary pattern · Hu invariant moments · Spatio-temporal space Multi-resolution analysis

---

M.N. Al-Berry (✉) · M.A.-M. Salem · H.M. Ebeid · M.F. Tolba
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
e-mail: maryam_nabil@cis.asu.edu.eg

M.A.-M. Salem
e-mail: salem@cis.asu.edu.eg

H.M. Ebeid
e-mail: halam@cis.asu.edu.eg

M.F. Tolba
e-mail: fahmytolba@gmail.com

A.S. Hussein
Faculty of Computing Studies, Arab Open University-Headquarters, 13033, Kuwait, Kuwait
e-mail: ashrafh@acm.org

T. Gaber et al. (eds.), *The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt.*,
Advances in Intelligent Systems and Computing 407, DOI 10.1007/978-3-319-26690-9_2

1 Introduction

Wavelets [1] and multi-scale techniques have been widely used in many image processing and computer vision applications [2, 3]. The most important property of wavelet analysis is the ability to analyze functions at different scales. Wavelets also provide a basis for describing signals accurately [4]. These properties of wavelet analysis have led to very successful applications within the fields of signal processing, image processing and computer vision. One of the fields that benefit from wavelets is the field of human action recognition [3, 5, 6].

In this way, wavelets have been used for spatio-temporal representation and description of human actions. In spatio-temporal methods, the action is represented by fusing spatial and temporal information into a single template. Spatio-temporal techniques can be divided into global and local techniques [7]. The advantages of both global and local representation techniques can be combined to obtain robust representations for human actions [8]. The 3D Wavelet Transforms [2, 6, 9] have been proposed and used to process video sequences as a 3D volume having time as the third dimension. The wavelet coefficients are then used to obtain global [3, 5, 10] or local [6] representations for the performed actions.

In this paper, a new method for action representation and description is proposed based on the wavelet analysis. The proposed method aims at extracting discriminative features and increasing the classification accuracy. The power of the 3D Stationary Wavelet Transform (SWT) is used to highlight spatio-temporal variations in the video volume and provides a multi-scale directional representation for human actions. These representations are described using a new proposed descriptor based on the concept of the Local Binary Pattern (LBP) [11], which is called the weighted directional wavelet LBP histogram. The proposed local descriptor is combined with Hu invariant moments [12] that are extracted from the Multi-scale Motion Energy Images (*MMEI*) [5] to get benefit from the global information contained in the wavelet representation. The performance of the proposed method is verified through several experiments using KTH dataset [13]. The rest of the paper is organized as follows: Sect. 2 reviews the background of the field and the state-of-the-art methods. The proposed representation and description method is illustrated in Sect. 3. In Sect. 4, the proposed method is applied, and the performance evaluation and comparison is performed. Finally, the work is concluded, and possible directions for improvement are pointed out in Sect. 5.

2 Related Work

The wavelet transforms [1, 14, 15] are being used in different fields such as signal, image and video compression as they provide the basis to describe signal features easily. The Decimated bi-orthogonal Wavelet Transform (DWT) [14] is probably the most widely used wavelet transform as it exhibits better properties than the

Continuous Wavelet Transform (CWT). It has much less scales than the CWT as only dyadic scales are considered. DWT is also very computationally efficient, since it requires O(N) operations for N samples of data and can be easily extended to any dimension by separable products of the scaling and wavelet functions [16].

When dealing with images in the spatial domain, two-dimensional DWT is used to decompose an image into four sub-images; an approximation and three details. The approximation A is obtained by applying the low-pass filter to the two spatial dimensions of the image and represents the image at a coarse resolution. The three detailed images are the horizontal detail H, the vertical detail V and the diagonal detail D [2]. Wavelet analysis has been used in the context of 2D spatio-temporal action recognition. For example, Siddiqi et al. [17] have used 2D DWT to extract features from the video sequence. The most important features were then selected using a Step Wise Linear Discriminant Analysis (SWLDA). Sharma et al. [18] have represented short actions using Motion History Images (MHI) [19]. They used two dimensional, 3 level dyadic wavelet transforms and modified their proposed representation to be invariant to translation and scale, but they concluded that the directional sub-bands alone were not efficient for action classification. In [20], Sharma and Kumar have described the histograms of MHIs by orthogonal Legendre moments and modeled the wavelet sub-bands by Generalized Gaussian Density (GGD). They have found that the directional information encapsulated in the wavelet sub-band enhances the classification accuracy.

Wavelet transforms have also 3D extension. Rapantzikos et al. have used the 3D wavelet transform to keep the computational complexity low while representing dynamic events. They have used wavelet coefficients to extract salient features. They treated the video sequence as a spatio-temporal volume, and local saliency measures are generated for each visual unit (voxel). Shao et al. [21] have applied a transform based technique to extract discriminative features from video sequences. They have shown that the wavelet transform gave promising results on action recognition.

While the decimated wavelet transform has been successful in many fields, it has the disadvantage of translation variance, so to overcome this drawback and maintain shift invariance, the Stationary Wavelet Transform (SWT) was developed. It has been proposed in the literature under different names such as; un-decimated wavelet transform [4], redundant wavelet [22] and shift-invariant wavelet transform [23]. The SWT gives a better approximation than the DWT since it is linear, redundant and shift-invariant [2]. In, the 3D SWT has been proposed and used for spatio-temporal motion detection. A 3D SWT is applied to each set of frames, and then the analysis is performed along the x-direction, the y-direction and the t-dimension of the time varying data [2], which is formed using the input frames.

Based on this 3D SWT, the authors of [10] have proposed two spatio-temporal human action representations. The proposed representations were combined with Hu invariant moments as features for action classification. Results obtained using the public dataset were promising and provide a good step towards better enhancements. They have proposed a directional global wavelet-based representation of natural human actions [3] that utilizes the 3D SWT [2] to encode the directional spatio-temporal characteristics of the motion.

Wavelets have also been used in combination with local descriptors such as Local Binary Patterns (LBPs) [11, 24] to describe texture images. LBPs are texture descriptors that have proved to be robust and computationally efficient in many applications [21, 25, 26]. The most important characteristic of LBPs is that it is not only computationally efficient, but also invariant to gray level changes caused by illumination variations [11]. The original LBP operator labels the pixels of an image by thresholding the 3×3 neighborhood of each pixel with the center value and considering the result as a binary number. The LBP was extended to operate on a circular neighborhood set of radius R. This produces 2^P different binary patterns for the P pixels constituting the neighborhood. Uniform LBP was defined as the pattern that contains at most 2 transitions for 1 to 0 or from 0 to 1. These patterns describe the essential patterns that can be found in texture. The disadvantage of LBP is lack of directional information.

In this paper, the 3D SWT is combined with LBPs and employed in the field of human action recognition. A new descriptor is proposed for describing human actions represented by the wavelet coefficients. The new descriptor combines the directional information contained in the wavelet coefficients in a weighted manner using the entropy value. The new local descriptor is expected to provide discriminative local features for the human actions.

3 Proposed Action Representation and Description

In this section, the proposed approach for extracting features is described. The proposed approach uses the power of the 3D SWT to detect multi-scale directional spatio-temporal changes and describes these changes using a weighted Local Binary Pattern (LBP) histogram. The weighted directional wavelet LBP is fused with moments to maintain global relationships. The proposed method is illustrated in Fig. 1.

3.1 Weighted Directional Wavelet Local Binary Pattern Histogram

The proposed descriptor is obtained by treating the video sequence as a 3D volume; considering time the third dimension. First, the video sequence is divided into disjoint blocks of frames and these blocks are supplied to the 3D SWT. The number of frames in one block depends on the intended number of wavelet analysis levels (resolutions). The number of frames is obtained by $L = 2^J$, where J is the number of resolutions. The 3D SWT produces the spatio-temporal detail coefficients

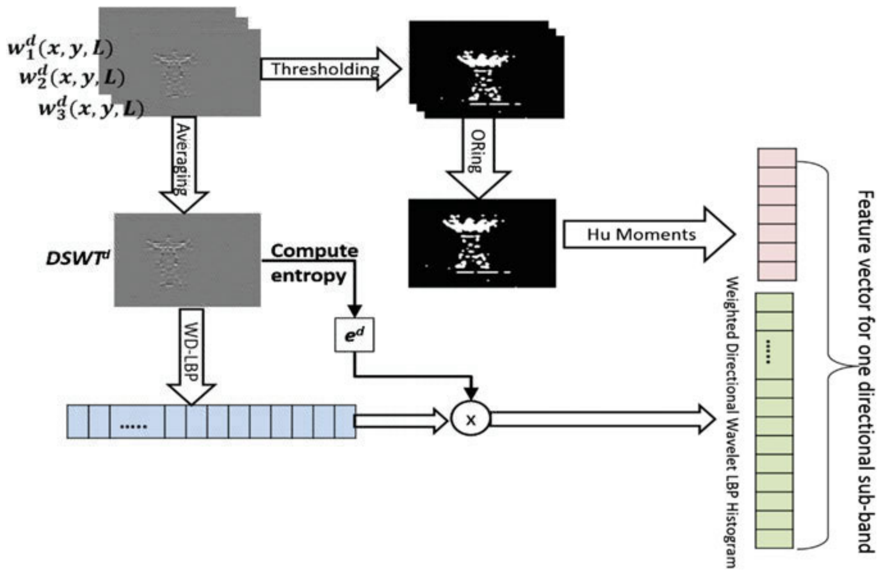


Fig. 1 Illustration of the proposed method

$w_j^d(x, y, t)$, where (x, y) are the spatial coordinates of the frames, t is the time, $j = 1, 2, \ldots, J$ is the resolution (scale) and d is the sub-band orientation, and the approximation coefficient $c_j(x, y, t)$ computed by the associated scaling function as described in [2].

The detail sub-bands $w_j^d(x, y, t), d = (5, 6, 7)$ are used for representing the action as they contain highlighted motion detected in the temporal changes that happened along the t-axis. Sub-band 4 also highlights temporal changes along with global spatial approximation. It is not used here because the focus is on investigating the effect of using local details only. The resulting coefficients are used to obtain a directional multi-scale stationary wavelet representation for the action by averaging the wavelet coefficients of the obtained resolutions at time $t = L$. This results in three Directional multi-scale Stationary Wavelet Templates $DSWT^d, d = (5, 6, 7)$, for the three detail sub-bands.

$$DSWT^d(x, y) = \frac{\sum_{j=1}^{3} w_j^d(x, y, L)}{\mathrm{J}}, d = (5, 6, 7) \tag{1}$$

Each directional stationary wavelet template is treated as texture images and the Local Binary Pattern is used to extract local features from it. The Directional Wavelet—LBP (DW-LBP) is computed as follows: