Engineering Applications of Artificial Intelligence 36 (2014) 54–63

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Measuring consistency of two datasets using fuzzy techniques and the concept of indiscernibility: Application to human perceptions on fabrics

Z. Xue [a], X. Zeng [a,*], L. Koehl [a], Y. Chen [b]

[a] GEMTEX, ENSAIT, 2 allée Louise et Victor Champier, BP30329, F-59056 Roubaix Cedex 1, France
[b] Department of Clothing Design and Engineering, College of Textile and Clothing Engineering, Soochow University, Suzhou, Jiangsu Province 215021, PR China

ARTICLE INFO

Article history:
Received 13 August 2013
Received in revised form
13 May 2014
Accepted 9 July 2014
Available online 7 August 2014

Keywords:
Consistency degree
Indiscernibility
Fuzzy techniques
Fabrics
Tactile properties
Visual representation

ABSTRACT

This paper presents an approach for developing a new consistency degree of two datasets, obtained from two different measuring systems on the same collection of items. In this approach, the concept of indiscernibility, frequently used in rough set approaches, is used to discover the classification consistency-based inclusion of one dataset to another. Next, in order to take into account the influence of neighboring relations of different data, we modify the previous index by proposing a fuzzy classification consistency-based inclusion degree. Also, the ordinal correlation between these two datasets, measured using a non-parametric method called Kendall's coefficient, is introduced. Finally, in order to create a reasonable integration of the previous two indices, a general consistency measure is constituted by introducing the expert knowledge into a fuzzy inference system. The overall procedure is believed to be capable of detecting nonlinear patterns lying beneath data while being safe to use a comparatively small number of experimental samples. Moreover, this new method can prevent the "black box" phenomenon encountered in many modeling techniques and produce robust and interpretable results. In practice, the proposed method is particularly significant for validating one measuring or evaluation system with respect to a standard reference. In order to validate the effectiveness of the proposed consistency degree, we apply it to study the relationship between tactile properties of a collection of fabric samples and their visual representations. The obtained results confirm that most of the tactile information can be perceived correctly by assessors through either video or image displays, while a better performance is detected in video scenarios.

1. Introduction

A great number of data mining methods have been developed for exploiting complex relations among multiple datasets. Nowadays, the most commonly used methods are based on statistics, including Linear Regression Analysis (Weisgerg, 2005), Principal Component Analysis (PCA) (Jolliffe, 2002), Multidimensional Scaling (Hollins et al., 1997; Picard, 2003), Multiple Factor Analysis (Howorth and Oliver, 1958; Le Dien, 2003) and various kinds of correlation coefficient analysis (Härdle and Simar, 2003). These methods are efficient in solving a lot of characterization and modeling problems from datasets due to their good capacity of identifying linear patterns from different information sources and then discovering correlations between data and between attributes from a big quantity of numerical data. And also for this reason, these methods have been widely employed to various research fields, including economics, medicine, biology, chemistry and engineering. (Agresti and Finlay, 1997).

However, modeling of relations between different datasets usually encounters problems of uncertainty and imprecision, and the classical methodologies are gradually showing their drawbacks in practice. First, when a problem is dealing with human knowledge, the concerned relations are often nonlinear. The application of the frequently used statistical techniques might cause important information loss due to their linearly structured models. Next, in many cases, there exists high uncertainty and imprecision in data structures due to non-unified linguistic human evaluation scores. But most of the classical analysis methods can only process perfect and complete numerical data without any uncertainty and imprecision. Next, the classical methods cannot always lead to precise and significant physical interpretation of data, and the obtained correlation results cannot be used to analyze all types of relations between datasets such as inclusion, causal and association relations. Finally, the classical methods often have

* Correspondence to: Laboratoire Génie et Matriaux Textiles (GEMTEX), École Nationale Supérieure des arts et Industries Textiles (ENSAIT), 2 allée Louise et Victor Champier, BP30329, F-59056 Roubaix Cedex 1, France. Tel.: +33 32025 89 67; fax: +33 32024 84 06.
E-mail address: Xianyi.zeng@ensait.fr (X. Zeng).

http://dx.doi.org/10.1016/j.engappai.2014.07.010
0952-1976/

strict requests on the size and distribution of the database. But collection of a great number of physically measured or human evaluation data is quite time-consuming and sometimes not practical for many research, for example in pilot studies. With a limited collection of samples, it is unlikely to obtain good fit modeling results using the classical methods.

In this situation, intelligent computational techniques, such as artificial neural network (ANNs) (Fausett, 1994), genetic algorithm (GA) (Goldberg, 1989), fuzzy logic (Zadeh, 1965; Sugeno and Yasukawa, 1993) and many hybrid applications of these tools (Ruan, 1997), have largely been applied to modeling and analysis with physical and human data. They have high capacity in (1) solving nonlinear problems, (2) dealing with both numerical and linguistic data, (3) modeling human expert reasoning so as to produce precise and straightforward interpretation of results, and (4) computing with relatively small sets of data and without need of any preliminary or additional information like probabilistic distributions in statistics. Of all these intelligent computational tools, the importance of the introduced notion of fuzzy set was properly realized by the research worker in all the branches of science and technology and has successfully been applied. Recently fuzzy set theory has been applied in matrix transformations by Tripathy and Baruah (2010), in mathematical analysis by Tripathy and Borgogain (2011), Tripathy et al. (2012), in topology by Tripathy and Debnath (2013), in mixed fuzzy topology by Tripathy and Ray (2012), in fuzzy logic by Sugeno and Yasukawa (1993), in rough mereology by Polkowski and Skowron (1994a) and many others.

In the current study, assuming that we have two datasets obtained from different measuring systems on one collection of items (products, consumers,…), we propose a novel method to measure how much the information quantity of one dataset is included in another one. It is particularly significant for validating one measuring or evaluation system with respect to a standard reference. For example, using this method, we can determine if a newly developed cheap measuring device can completely take into account the main features given by an old expensive measuring system. The proposed method was built using fuzzy techniques and the concept of indiscernibility of rough sets theories (Pawlak, 1982).

Rough set theory has been widely applied for measuring classification-based consistency or inclusion degree of two datasets (Qian et al., 2008; Xu et al., 2012). In practice, it is considered as more relevant for processing small sets of objects. However, in our study, there exist two following particularities different from the classical application scenarios of rough set theory: (1) the objects in our datasets can be fully arranged in order for each specific attribute while those in a general dataset can just be classified and (2) the objects in our datasets mostly correspond to real measuring data. Therefore, in this special context, the classical inclusion degree, dealing with classification of objects, has been modified by introducing fuzzy techniques in order to take into account of neighboring relations of objects and their relations with respect to normalized measuring scales. The crisp partition of objects is then transformed into a fuzzy partition of objects related to a set of normalized scores. Also, the general consistency has been obtained by combining the previous classification-based consistency and Kendall's coefficient-based ranking consistency. Therefore, the equivalence and neighboring relations of objects as well as ranks of objects in each dataset have been taken into account in this general consistency degree. The overall consistency criterion is believed to be capable of detecting nonlinear patterns lying beneath datasets while being safe to use a comparatively small number of experimental samples. Moreover, this new method can prevent the "black box" phenomenon encountered in many modeling techniques and produce robust and interpretable results.

2. Fuzzy classification consistency of two datasets

2.1. Classification consistency (CCons)

The classification consistency-based (*CCons*) inclusion degree of one dataset to another one is defined by using the concept of indiscernibility of the rough sets theory, developed by Zdzislaw Pawlak in the early 1980s. It is a relatively new soft computing tool for solving imperfect data analysis. Rough set approach has become a popular mathematical framework in many research areas such as data mining, knowledge discovery from database, decision support, feature selection and pattern recognition (Pawlak, 1998; Pawlak and Skowron, 2007; Qian et al., 2007). According to Pawlak (1998), *indiscernibility relation*, characterizing equivalence relations of different objects, is a central concept of rough set theory. In our paper, this concept, together with the other concepts frequently used in rough set theory, such as *decision table* and *rough mereology*, constitute the main theoretical foundation of the proposed approach.

2.1.1. Indiscernibility

A basic assumption in rough set theory is that every object of the universe of discourse is associated with some information (data, knowledge). Objects characterized by the same information are *indiscernible* (*similar*) with the available knowledge about them. The *indiscernibility relation* generated in this way is the mathematical basis of rough set theory. Any set of all indiscernible objects is called an elementary set, and expresses a basic granule of knowledge about the universe of discourse. Any union of some elementary sets is referred to as a crisp (precise) set, otherwise the set is rough (imprecise). In other words, a crisp set expresses knowledge that can be subdivided into all discernible granules, whereas a rough set cannot be precisely characterized with available knowledge, or contains knowledge that cannot be decisively subdivided or discerned. Since rough sets theory addresses granularity of knowledge, expressed by the indiscernibility relation, we are unable to deal with single object but we have to consider clusters of indiscernible objects, as fundamental concepts of the theory.

2.1.2. Information system and decision table

From the above remarks, we know that rough set theory in fact deals with the classificatory analysis of information (data, knowledge). A rough set-based data analysis starts from a data table, called an information system.

Formally, an information system is a pair $S=(U, A)$, where

(a) U is a non-empty finite set of objects, or the universe;
(b) A is a non-empty finite set of attributes; and
(c) for every $a \in A$, there is a mapping $a: U \to V_a$, where V_a is called the value set of a.

Any subset of attributes $P \subseteq A$ determines a binary *indiscernibility relation* $IND(P)$ defined by

$$IND(P) = \{(u, v) \in U \times U \mid \forall a \in P,\ a(u) = a(v)\}$$

Obviously, $IND(P)$ is an equivalence relation on the set U. For $P \subseteq A$, the relation $IND(P)$ constitutes a partition of U, which is denoted by $U/IND(P)$, or just U/P. An information system in which values of all attributes for all objects from U are known is called complete, otherwise it is incomplete.

If an information system is distinguished into two disjoint classes of attributes, named condition and decision attributes, respectively, then the system will be called a decision table and will be denoted by $S=(U, C \cup D)$, in which C and D are disjoint sets of condition and decision attributes, respectively. In the same