Information Systems 69 (2017) 124–154

Contents lists available at ScienceDirect

Information Systems

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

Dispersed decision-making system with fusion methods from the rank level and the measurement level – A comparative study

Małgorzata Przybyła-Kasperek*, Alicja Wakulicz-Deja

University of Silesia, Institute of Computer Science, Będzińska 39, Sosnowiec 41-200, Poland

ARTICLE INFO

Article history:
Received 23 September 2016
Revised 15 February 2017
Accepted 8 May 2017
Available online 11 May 2017

Keywords:
Decision support system
Dispersed knowledge
Conflict analysis
Fusion method
Combining classifiers

ABSTRACT

This article discusses the problem of decision-making based on dispersed knowledge that is stored in several independent knowledge bases. The dispersed decision-making system, which was proposed in a previous paper of the authors, is used. In this study, four fusion methods from the rank level and nine methods from the measurement level were used in this dispersed system. These methods were tested on three data sets from the UCI Repository – Soybean, Vehicle Silhouettes and Landsat Satellite. The sets are diverse in terms of the number of objects, the number of conditional attributes and the number of decision classes. There are also various types of conditional attributes in these sets. The experimental section is divided according to the three objectives of the article. The fusion methods were compared in the two groups – rank and measurement levels. In addition, experiments were carried out fusing multiple methods simultaneously in the decision-making process. Methods from the rank level and the measurement level were applied simultaneously in the same decision-making process. Then the decisions that were generated by the methods were merged. The results were compared and conclusions were drawn. The third goal of the article was to compare the efficiency of the inference of fusion method with and without the use of a dispersed system. It was found that the use of a dispersed system improved the efficiency of inference in most cases.

1. Introduction

Methods of inference were created in order to be used when we have one knowledge base and when we want to make a decision. Over time, a more complex problem arose that involved the simultaneous use of various methods of inference in order to improve the efficiency of inference. Moreover, the approach of dividing one knowledge base with respect to features or objects and then inferring based on these smaller bases in order to achieve greater efficiency was considered. In such situations, it usually happened that when decisions were merged there were differences in the decisions taken by the various classifiers. A variety of methods, whose aims were shelling the strengths of classifiers and avoiding their weaknesses so as to achieve improvements in efficiency, were proposed.

In this study fusion methods were used in a different situation. We assumed that a certain set of knowledge bases is given in advance. These databases can be collected separately. They may have different features and objects. Only the field to which they relate, namely the decision attribute, is shared across all of the databases. The concept of using dispersed knowledge has been considered by the authors for several years. Various approaches were considered – the static approach [24,43] and the dynamic approach [25,26,29,30]. In the paper [26], it was justified that the proposed dispersed decision-making system with a dynamically generated structure generates the best results. Therefore, in this study, this approach was used in conjunction with various fusion methods.

Different methods of combining classifiers are used depending on what information we obtain from the base classifiers. In the literature [14,45], three types of classifier outputs are distinguished. Type 1 – the abstract level in which each base classifier generates a name or a number of the class to which the observation belongs. Type 2 – the rank level in which each base classifier generates a set of classes ordered by the plausibility that they are the correct labels. Type 3 – the measurement level in which each base classifier generates a vector that represents the probability of an observation belonging to different decision classes.

In this study, four fusion methods from the rank level and nine methods from the measurement level were used in a dispersed decision-making system. The following methods from the rank level were examined – the Borda count, the intersection, the

* Corresponding author.
E-mail addresses: malgorzata.przybyla-kasperek@us.edu.pl, alicja.wakulicz-deja@us.edu.pl (M. Przybyła-Kasperek), alicja.wakulicz-deja@us.edu.pl (A. Wakulicz-Deja).

http://dx.doi.org/10.1016/j.is.2017.05.002
0306-4379/

highest rank and the union method. These methods belong to two groups – the group of class set reordering methods (the Borda count and the highest rank method) and the group of class set reduction methods (the intersection and the union method). Nine different methods from the measurement level were considered. The simplest methods from this group are the maximum rule, the minimum rule, the median rule, the sum rule and the product rule. More complex methods such as the weighted average method and the probabilistic product method were also considered. The most sophisticated methods that were analysed were the method using decision templates and the method that is based on the theory of evidence. In addition, four different distance measures were analysed for the method using decision templates: the similarity measure that uses the normalised Euclidean distance, the similarity measure that uses the symmetric difference defined by the Hamming distance, the Jaccard similarity coefficient and the similarity measure that uses the symmetric difference.

The concept of applying the fusion methods in a dispersed decision-making system was considered in the papers [27,28] in a fragmentary way. Those articles only described preliminary experiments and a comprehensive comparison of methods was not provided in them. Moreover, the approach without the use of a dispersed system and the approach using multiple methods in one decision-making process were not considered in those papers. These issues are addressed in this paper. In addition, in the experimental section of this article, an additional set of data, which is the largest of the three data sets used, is considered and its analysis is an important contribution. The main contributions of this paper are listed below:

- performing a comprehensive comparison of the fusion methods in the two groups – the rank and the measurement levels,
- performing a comparison of the approaches with and without a dispersed decision-making system,
- performing a test of the approach using several methods in one decision-making process,
- conducting experiments using an additional set of data, which is the largest of the three data sets used.

The paper is organised as follows. The second section presents the related papers. The third section briefly describes the dispersed decision-making system. The fourth section describes the fusion methods that were used. The fifth section presents an example of the use of fusion methods in a dispersed system. The sixth section shows a description and the results of experiments carried out using three data sets from the UCI repository: Soybean, Vehicle Silhouettes and Landsat Satellite are used. These data sets are diverse in terms of the number of objects, the number of conditional attributes and the number of decision classes. This section is divided into three parts. The first part of this section contains the results of experiments using a dispersed system and fusion methods. The second part contains the results of experiments with the approach in which many fusion methods from the same level is used simultaneously in one decision-making process. The last part of the third section presents the results of experiments without using a dispersed system. The article concludes with a short summary in the seventh section.

2. Related work

The issue of combining classifiers is a very important aspect in the literature [12,14]. Classifier ensembles were used in different situations [11,15,23]. The aim of the issue is always to improve the quality of the classification by combining the results of the predictions of the base classifiers. One of the basic questions is what combination rule to use. In this article different fusion methods are considered. These methods are very popular and are described in numerous papers [4,12,14,41,45]. In this paper various applications of the combination rule are considered - methods are used individually as well as several of them are used simultaneously during one decision-making process.

In this article we consider a set of classifiers that make decisions based on dispersed knowledge. The issue of making decisions based on dispersed knowledge is widely discussed in the literature. For example, this issue is discussed in the multiple model approach [4,14,22]. In a multiple classifier system, an ensemble is constructed that is based on base classifiers. The aim of this approach is to reduce any misclassification at the cost of increased computational complexity. Ensemble accuracy depends on both the quality of the problem decomposition and the individual accuracies in the base classifiers. One of the methods for decomposition is to use the domain knowledge to decompose the nature of the decisions into a hierarchy of layers [17]. In the papers [19,37,44], an ensemble of feature subsets is considered. The paper [38] investigates a correspondence between mathematical criteria for feature selection and mathematical criteria for voting between the resulting decision models. In the paper [8], a random subspace technique for building an ensemble is considered. A very important matter is that some form of diversity among the base classifiers must exist in order to improve accuracy [35,40]. The method for generating the final decision also has a significant impact on the efficiency of the ensemble [3].

In the book [5], an overview of the various aggregation methods for different types of data is included. Among others, the concept of a general fusion function hierarchy using a neural network is considered. The book [42], provides a broad overview of the aggregation operators that are applied in the synthesis of judgements and information fusion. The aggregation operator is a function, that takes N numerical values and returns another numerical value. The paper [33], considers a situation in which agents use an adaptive judgement for risk assessment, risk treatment and cost/benefit analysis. In this approach, some hierarchy occurs and the aggregation of information (decision) systems with respect to some constraints is performed. Information granules are constructed at different levels of this hierarchy. The importance of hierarchical modelling and aggregation for many real-life projects is also emphasised in the papers [1,17]. Based on the approaches mentioned above, a need for hierarchical aggregation appears to be of special importance. In this paper, aggregation at different levels of hierarchy are presented. On the lower level of the hierarchy, the aggregation of decision tables is performed, while at the higher level, decisions are aggregated using fusion methods.

In the paper [6], it is noted that the domain knowledge can be crucial for improving efficiency in solving real-world problems. The domain knowledge should not only be expressed in some control rules, but also some additional explanation should be given in the system. In the paper [32], the process models from data and domain knowledge within the program Wisdom technology is discussed. The aim of the approach is to achieve the ability to make judgments correctly to a satisfactory degree, having in mind real-life constraints. In the paper [33], the importance of domain knowledge for dealing with Big Data in real-life applications is underlined. In order to model interactive computations, performed by the agent in complex systems based on Big Data, the complex granules concept is introduced. Some constraints and dependencies between objects can be defined by domain ontology. In the paper [9], it is noted that the use of domain knowledge, represented by the ontology of concepts, is very useful in real-life applications. In the paper [18], the domain knowledge is used to solve the problem of sunspot classification from satellite images. The domain knowledge is represented by the ontology of concepts – a treelike structure that describes the relationship between the target concepts, intermediate concepts and attributes is used. As can be seen the