Ensemble Evidential Editing k-NNs through rough set reducts

Asma Trabelsi

Université de Tunis, Institut Supérieur de Gestion de Tunis, LARODEC , Tunisia
Univ. Artois, EA 3926, Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A), Béthune, F-62400, France
**E-mail: trabelsyasma@gmail.com*

Zied Elouedi

Université de Tunis, Institut Supérieur de Gestion de Tunis, LARODEC , Tunisia
**E-mail: zied.elouedi@gmx.fr*

Eric Lefèvre

Univ. Artois, EA 3926, Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A), Béthune, F-62400, France
**E-mail: eric.lefevre@univ-artois.fr*

Ensemble classifier is one among the machine learning hot topics and it has been successfully applied in many practical applications. Since the construction of an optimal ensemble remains an open and complex problem, several heuristics for constructing good ensembles have been introduced for several years now. One alternative consists of integrating rough set reducts into ensemble systems. To the best of our knowledge, almost existing methods neglect knowledge imperfection, knowing that several real world databases suffer from some kinds of uncertainty and incompleteness. In this paper, we develop an ensemble Evidential Editing k-Nearest Neighbors classfier (EEk-NN) through rough set reducts for addressing data with evidential attributes. Experimentations in some real databases have been carried out with the aim of comparing our proposal to another existing approach.

Keywords: Ensemble classifiers, rough set reducts, Evidential Editing k-Nearest Neighbors classfier, evidence theory.

1. Introduction

Ensemble system has attracted a great attention since 1990s thanks to its prediction performance ability.[4] Diversity between classifiers represents a key element for designing good successful ensembles.[11] Manipulating the input feature space has been theoretically and experimentally defined as a sufficient way for establishing high diversity between base classifiers.[1,6,9,24] The choice of the most suitable feature subsets for constructing ensemble systems is still an open question. Recently, feature subsets yielded through rough set reducts[14] have been successfully introduced into ensemble systems.[15,17,18,25] It must be emphasized that almost all real world data are vulnerable to incompleteness, inconsistency and imprecision. This imperfection may pervade either the attribute values, the class labels or both of them. Despite its importance, little attention has been drawn to extract reducts from a such kind of data. In this paper, we are only interested to data with uncertain attribute values represented within the evidence theory[21] and we aim to construct an ensemble of the Evidential Editing k-Nearest Neighbor classifier (EEk-NN)[23] through rough set reducts to process uncertainty. The remaining of this paper is organized as follows: Section 2 is dedicated to recall some basic concepts of the evidence theory. We describe, in Section 3, our novel ensemble system framework. We present, in Section 4, our experimentations on several synthetic databases. Finally, the conclusion and our main future work directions are reported in Section 5.

2. Basic concepts of the evidence theory

The frame of discernment Θ constitutes a finite non empty set of elementary hypotheses.[16] An expert's belief over a given subset of Θ has to be represented by the so-called basic belief assignment m (bba) fulfilling:

$$\sum_{A \subseteq \Theta} m(A) = 1 \tag{1}$$

The simple support function (ssf) is a special case of the basic belief assignments. It has two focal elements: the frame of discernment Θ and a strict subset of Θ which named the focus of the ssf.[19]

The evidence theory provides a set of combination rules for merging distinct information sources. Dempster's rule is one of the best known rules. Given two information sources S_1 and S_2 with respectively m_1 and m_2 as bbas, Dempster's rule,[3] denoted by $\oplus$, will be set as:

$$m_1 \oplus m_2(A) = \frac{1}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)} \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Theta \tag{2}$$

3. Classifier ensemble through rough set reducts

In this paper, we present a new classifier ensemble framework for processing imperfect knowledge. More concretely, we propose an ensemble of our EEk-NN classifier[23] through rough set reducts for handling data described by evidential attributes. The proposed framework is detailed in Algorithm 3.1. It consists of two main levels. The first one concerns the generation of reducts from a given uncertain data, while the second one selects reducts enabling the construction of a successful EEk-NN ensemble. We present in what follows each of these steps.

3.1. *A novel framework for generating reducts from uncertain data*

A number of solutions have been proposed for dealing with multiple reduct generation problems. The Rosetta software is well known to be among the most effective alternative.[10] It provides a set of algorithms for multiple reduct extraction. An example includes the SAVGenetic Reducer that implements a genetic algorithm for searching approximate hitting sets, meaning approximate reducts.[5] One limitation of this latter is its inability to process uncertainty. In this paper, we propose an extension of the SAVGenetic algorithm for addressing data with uncertain attribute values that are expressed in terms of evidence. In accordance with the standard SAVGenetic reducer, our proposal starts by computing a discernability matrix from a given data. We have developed in a previous work,[21] a novel algorithm allowing the computation of a belief discernability matrix Λ' from data with evidential attributes. Let O=$\{O_1$,..,$O_N\}$ be a given data described by a finite non empty set of N objects. Each object i ($i \in \{1, \ldots, N\}$) is defined by a set of n uncertain attributes $uA = \{A_1, \ldots, A_n\}$ with values $uV^i = \{uv_1^i, \ldots, uv_n^i\}$ and a certain class label $Y_i \in C = \{c_1, \ldots, c_Q\}$. Suppose that Θ_k denotes the frame of discernment of the attribute A_k ($k \in \{1, \ldots, n\}$). Every uncertain attribute value uv_k^i of an instance O_i is represented by a basic belief assignment $m_i^{\Theta_k}$. Assume that S refers to a tolerance threshold (i.e. S is set to 0.1 with the aim of maximizing the search space) and $dist$ reflects the Jousselme distance.[8] The entries of the belief discernibility matrix Λ' are computed as follows:

$$\Lambda'(O_i, O_j) = \{A_k \in uA | Jousselme\_Dist(m_i^{\Theta_k}, m_j^{\Theta_k}) > S \text{ and } Y_i \neq Y_j\} \quad (3)$$

The non empty set of Λ' will then be stored in a multiset ζ'. The approximate hitting sets of ζ' correspond to the approximate reducts. For picking out the approximate hitting sets, we relied on the genetic algorithm with the following fitness function for each subset $B \in 2^n$:

$$f(B) = (1-\alpha) \times \frac{|A| - |B|}{|B|} + \alpha \times min\{\varepsilon, \frac{|[F \in \zeta' | F \cap B \neq \emptyset]|}{|\zeta'|}\} \qquad (4)$$

The fitness function $f(B)$ consists mainly on two terms. The former one rewards subsets with shortest size and the latter one rewards subsets that are hitting sets (i.e. meaning subsets having a non empty intersection with all elements of the discernability matrix). Herein, $\alpha \in [0,1]$ refers to the adaptive weighting between the two parts and ε reflects the minimal hitting set fraction.

3.2. *Reduct selection for ensemble learning*

An ensemble system with rough set reducts has been viewed for some years as a valid alternative for getting optimal performance.[7] Since several reducts may be generated for a given data set, the choice of the appropriate ones remains a field of research to further develop. Herein, we draw our inspiration from a study conducted in[13] for finding out the suitable reducts for an ensemble of EEk-NN classifiers when relied on both the accuracy and the diversity of base classifers. That is an appropriate trade-off between the diversity of classifiers and the accuracy of each individual classifier is really sufficient for yielding good performance. The assessment function that balances the accuracy and the diversity of base classifiers is as follows:

$$Fitness(f, L) = Accuracy(f, L) + \omega \times Diversity(f, L) \qquad (5)$$

where L is the number of classifiers, $Accuracy(f, L)$ reflects the average accuracy of the base classifiers, $Diversity(f, L)$ represents the diversity between base classifiers and ω corresponds to the parameter that balances Accuracy and Diversity. It is worth noting that there are several classifier diversity measures. Authors in[11] have distinguished pairwise and non-pairwise diversity measures. The choice of the most convenient one remains unanswered question. In this paper, we relied on the disagreement measure, which is a pairwise one, for computing classifier diversity. Concerning the parameter ω, it has to be adjusted automatically for maximizing the fitness function value.[13] In addition to the accuracy and diversity of the base classifiers, ensuring diversity between reducts has also been regarded as a substantial key element when designing ensemble systems. In fact, we aim to reduce the searching space of reducts by taking into consideration the diversity measure proposed in.[2] It is set to:

$$Div_{R_k} = 1 - \frac{\frac{|R_k \cap Selected\_Red|}{|R_k \cup Selected\_Red|}}{NB\_Selected\_Reduct} \tag{6}$$

where R_k is the candidate reduct, $Selected\_Red$ reflects the selected reducts and $NB\_Selected\_Reduct$ states the number of selected reducts. The candidate reducts with a diversity measure smaller than a threshold T will then be removed from the search space.

Algorithm 3.1 Successful rough set ensemble framework

Input: An uncertain data, M is the maximum chosen reducts.
Output: ensemble system.
% Subsection 3.1
Find multiple reducts *Reducts*
% Subsection 3.2
$Selected\_Red \leftarrow \emptyset$, $Ens\_Classifier \leftarrow \emptyset$
Choose the reduct R_1 with the lowest weight from the reduct pool *Reducts*, $Selected\_Red \leftarrow \{Selected\_Red, R_1\}$, $NB\_Selected\_Reduct \leftarrow 1$
$Reducts \leftarrow Reducts\text{-}R_1$
$Ens\_Classifier \leftarrow \{Ensemble\_Class,\ f_1\}$
do
Compute the diversity between $R_k \in Reducts$ and $Selected\_Red$
$Reduct\_To\_Remove \leftarrow$ all $R_k \in Reducts$ fulfilling $Div_{R_k} < T$
$Reducts \leftarrow Reducts$ - $Reduct\_To\_Remove$
Choose a new reduct R_j from *Reducts* satisfying:
$Fitness(f_j, Ens\_Classifier) = \max_{R_k \in Reducts}(Fitness(f_k, Ens\_Classifier))$
$Ens\_Classifier \leftarrow \{Ensemble\_Class,\ f_j\}$, $Selected\_Red \leftarrow \{Selected\_Red, R_j\}$, $NB\_Selected\_Reduct \leftarrow NB\_Selected\_Reduct{+}1$
until $NB\_Selected\_Reduct = M$ **or** isempty(*Reducts*)=true
Ensemble system merged through the Dempster operator

4. Experimentations settings and results

Throughout this paper, we propose to construct an ensemble of EEk-NN classifiers from data with evidential attributes. Since real world applications suffer from incompleteness and uncertainty, there is a lack of datasets that take imperfection into consideration. With the aim of evaluation of proposed approach, we propose to generate synthetic databases. The underling idea consists of injecting an uncertainty level P to some real categorical databases delivered by the the UCI machine learning repository.[12] Table 1 describes the used databases for experimentations. Getting inspiration from the method proposed in,[20] four uncertainty levels P have been considered:

an certain case when P=0, a Low uncertainty case when $(0 < P < 0.4)$, a Middle uncertainty case when $(0.4 \leq P < 0.7)$ and a High uncertainty case $(0.7 \leq P \leq 1)$. So that, each attribute value has to be expressed by a simple support mass function, meaning P has to be assigned be the focus reflecting the true attribute value and $1-P$ has to be allocated to the frame of discernment of that attribute.

Table 1. Description of databases

Databases	#Instances	#Attributes	#Classes
Voting Records	435	16	2
Monks	432	7	2
Lymphography	148	18	4
Tic-Tac-Toa	958	9	2

Our ensemble EEk-NN classifiers through rough set reducts is evaluated and compared to an ensemble of 25 EEk-NN classifiers through Random Subspaces presented in[22] and we have followed a 10-fold cross validation approach. Taking k=3 as nearest neignbors, the obtained Percentage of Correct Classifications (PCCs) are presented in Table 2 where ERR and ERS reflect respectively the ensemble EEk-NN through rough set reducts and the ensemble EEk-NN through random subspaces and $size$ represents the size of an ensemble constructed using the ERR approach.

Table 2. PCCs results

	No		Low		Middle		High	
	ERS	ERR	ERS	ERR	ERS	ERR	ERS	ERR
Voting Records	91.62	**93.02**	91.92	**95.35**	91.39	**95.12**	89.53	**90.11**
Monks	60.26	**100**	59.49	**100**	60.26	**100**	53.68	**68.18**
Lymphography	82.85	**87.90**	75.14	**79.29**	82.85	**84.12**	62.85	**75.66**
Tic-Tac-Toa	61.15	**72.53**	55.78	**59.05**	56	**74.95**	57.68	**57.76**

The obtained results have proven the performance of the rough set reduct method over the random subspace approach. In fact, the yielded PCCs through the rough set techniques are strictly higher than those obtained using the random subspace method. Let us take the Monks database with uncertainty equals $High$ as an example. The PCC derived by ensemble rough sets is equal to 68.18 %, while that achieved by the random subspace method equals 53.68 %.

5. Conclusion

We have proposed a novel framework for classifier ensemble for addressing data with evidential attribute values. Precisely, we have developed an ensemble of the EEk-NN classifier by relying on some rough set techniques for generating suitable feature subsets. For the purpose of assessing our novel approach, we have made a comparative study with an ensemble EEk-NN constructed via random subspaces. The achieved PCC results have proven the performance of our novel framework over that generated with random subspaces. Although, there are several combination operators within the evidence theory, in this paper we have merged classifier using the Dempster rule as it is very well known, in future work, we look forward to paying more attention to the combination procedure. Notably, we intend to pick out the best combination rule within the context of ensemble evidential classifiers.

References

1. R. Bryll, R. Gutierrez-Osuna, and F. Quek. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition*, 36(6):1291–1302, 2003.
2. E. Debie, K. Shafi, C. Lokan, and K. Merrick. Reduct based ensemble of learning classifier system for real valued classification problems. In *IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL)*, pages 66–73. IEEE, 2013.
3. A. P. Dempster. Upper and Lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.
4. T. G. Dietterich. Machine-learning research. *AI magazine*, 18(4):97, 1997.
5. M. A. El-Monsef, M. Seddeek, and T. Medhat. Classification of sand samples according to radioactivity content by the use of euclidean and rough sets techniques. *EG EG0600123*, page 266, 2003.
6. S. Günter and H. Bunke. Feature selection algorithms for the generation of multiple classifier systems and their application to handwritten word recognition. *Pattern recognition letters*, 25(11):1323–1336, 2004.
7. Y. Guo, L. Jiao, S. Wang, S. Wang, F. Liu, K. Rong, and T. Xiong. A novel dynamic rough subspace based selective ensemble. *Pattern Recognition*, 48(5):1638–1652, 2015.
8. A. Jousselme, D. Grenier, and E. Bossé. A new distance between two bodies of evidence. *Information fusion*, 2(2):91–101, 2001.
9. Y. Kim. Toward a successful crm: variable selection, sampling, and ensemble. *Decision Support Systems*, 41(2):542–553, 2006.
10. J. Komorowski, A. Øhrn, and A. Skowron. The rosetta rough set software system. *Handbook of data mining and knowledge discovery*, pages 2–3, 2002.
11. L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensem-

bles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.
12. C. K. Murphy. Combining belief functions when evidence conflicts. *Decision support systems*, 29(1):1–9, 2000.
13. D. W. Opitz. Feature selection for ensembles. *AAAI/IAAI*, 379:384, 1999.
14. Z. Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 11(5):341–356, 1982.
15. S. Saha, C. Murthy, and S. K. Pal. Classification of web services using tensor space model and rough ensemble classifier. In *International Symposium on Methodologies for Intelligent Systems*, pages 508–513. Springer, 2008.
16. G. Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.
17. L. Shi, X. Ma, L. Xi, Q. Duan, and J. Zhao. Rough set and ensemble learning based semi-supervised algorithm for text classification. *Expert Systems with Applications*, 38(5):6300–6306, 2011.
18. L. Shi, L. Xi, X. Ma, M. Weng, and X. Hu. A novel ensemble algorithm for biomedical classification based on ant colony optimization. *Applied Soft Computing*, 11(8):5674–5683, 2011.
19. P. Smets. The canonical decomposition of a weighted belief. In *IJCAI*, volume 95, pages 1896–1901, 1995.
20. N. Sutton-Charani, S. Destercke, and T. Denoeux. Learning decision trees from uncertain data with an evidential EM approach. In *Proceedibgs of the 12th International Conference on Machine Learning and Applications, ICMLA 2013, 2013, Volume 1*, pages 111–116, 2013.
21. A. Trabelsi, Z. Elouedi, and E. Lefevre. Feature selection from partially uncertain data within the belief function framework. In *proceesings of the 16th International Conferenc Information Processing and Management of Uncertainty in Knowledge-Based IPMU, Part II*, pages 643–655, 2016.
22. A. Trabelsi, Z. Elouedi, and E. Lefevre. Ensemble enhanced evidential k-nn classifier through random subspaces. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 212–221. Springer, 2017.
23. A. Trabelsi, Z. Elouedi, and E. Lefevre. A novel k-nn approach for data with uncertain attribute values. In *Proceedings of the 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Part I*, pages 160–170, 2017.
24. K. Tumer and N. C. Oza. Input decimated ensembles. *Pattern Analysis & Applications*, 6(1):65–77, 2003.
25. S.-L. Wang, X. Li, S. Zhang, J. Gui, and D.-S. Huang. Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction. *Computers in Biology and Medicine*, 40(2):179–189, 2010.