Knowledge-Based Systems 88 (2015) 57–69

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

EK-NNclus: A clustering procedure based on the evidential K-nearest neighbor rule

Thierry Denœux [a,*], Orakanya Kanjanatarakul [a,b], Songsak Sriboonchitta [c]

[a] Sorbonne Universités, Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, France
[b] Faculty of Management Sciences, Chiang Mai Rajabhat University, Thailand
[c] Faculty of Economics, Chiang Mai University, Thailand

ARTICLE INFO

Article history:
Received 2 May 2015
Received in revised form 24 July 2015
Accepted 11 August 2015
Available online 20 August 2015

Keywords:
Dempster–Shafer theory
Evidence theory
Hopfield neural networks
Unsupervised learning
Credal partition

ABSTRACT

We propose a new clustering algorithm based on the evidential K nearest-neighbor (EK-NN) rule. Starting from an initial partition, the algorithm, called EK-NNclus, iteratively reassigns objects to clusters using the EK-NN rule, until a stable partition is obtained. After convergence, the cluster membership of each object is described by a Dempster–Shafer mass function assigning a mass to each cluster and to the whole set of clusters. The mass assigned to the set of clusters can be used to identify outliers. The method can be implemented in a competitive Hopfield neural network, whose energy function is related to the plausibility of the partition. The procedure can thus be seen as searching for the most plausible partition of the data. The EK-NNclus algorithm can be set up to depend on two parameters, the number K of neighbors and a scale parameter, which can be fixed using simple heuristics. The number of clusters does not need to be determined in advance. Numerical experiments with a variety of datasets show that the method generally performs better than density-based and model-based procedures for finding a partition with an unknown number of clusters.

1. Introduction

Clustering may be defined as the search for groups in data, in an unsupervised way. Over the years, the initial concepts of hierarchical and partitional clustering have been extended to the search for more complex data structures, leading to the notions of fuzzy [2], possibilistic [19], rough [22], or credal clustering [7,27]. In spite of the huge amount of work done in this area, the design of computationally efficient algorithms able to reveal an informative structure in data remains a topic of considerable interest.

In the so-called "decision-directed" approach to clustering [page 536] [10], prior knowledge is used to design a classifier, which is used to label the samples. The classifier is then updated, and the process is repeated until no changes occur in the labels. For instance, the well-known c-means algorithm is based on this principle: here, the nearest-prototype classifier is used to label the samples, and it is updated by taking as prototypes the centers of each cluster. The classification EM algorithm [4] is also based on the same principle, with an arbitrary parametric classifier and maximum likelihood estimation.

In recent years, new approaches to classification and clustering using the Dempster–Shafer theory of belief functions [5,31] have been developed. For supervised classification, one of the most widely used method is the evidential K-nearest neighbor (EK-NN) rule [6,42]. Variants of this method have been proposed in [21,23,24,30,41]. This approach has been successfully applied in many domains including bioinformatics [32,33,38,39], medical image processing [3], remote sensing [41], machine diagnosis [37], process control [34], among others. In unsupervised learning, the notion of credal partition has been introduced in [7]. In a credal partition, the member of an object to clusters is described by a Dempster–Shafer mass function. The ECM algorithm, a c-means-like algorithm that generates credal partitions, was introduced in [27]. Variants of this algorithm were proposed in [20,26,40].

In this paper, we introduce a new decision-directed evidential clustering algorithm based on the EK-NN rule. Starting from an initial random partition, the label of each sample is updated in turn using the EK-NN rule. We prove that this algorithm, (called Ek-NNclus) converges to a fixed point that corresponds, under some assumptions, to the most plausible partition of the data. We also show that the algorithm can be implemented in a competitive Hopfield neural network [15,16], which allows its possible parallelization. The method is simple and depends on a small number

* Corresponding author. Tel.: +33 344 234 496; fax: +33 344234477.
E-mail address: tdenoeux@utc.fr (T. Denœux).

http://dx.doi.org/10.1016/j.knosys.2015.08.007
0950-7051/

of easily tunable parameters. In particular, it does not require to fix the number of clusters in advance. After convergence, one obtains a credal partition, which is more informative than a fuzzy partition and allows us to easily detect outliers.

The rest of this paper is organized as follows. The background on belief functions, the EKNN rule and credal clustering will first be recalled in Section 2. The new clustering algorithm will then be introduced and its theoretical properties will be studied in Section 3. Finally, experiments will be presented in Sections 4 and 5 will conclude the paper.

2. Background

This section is intended to make the paper self-contained, by recalling the necessary concepts related to belief functions (Section 2.1), the EK-NN rule (Section 2.2) and credal clustering (Section 2.3).

2.1. Belief functions

The theory of belief functions (also referred to as Dempster–Shafer, or evidence theory) is a framework for reasoning under uncertainty based on the explicit representation and combination of items evidence [5,31]. Let us assume that we are interested in the value of some variable $\boldsymbol{\omega}$ taking values in a finite domain Ω, called the *frame of discernment*. Uncertain evidence about $\boldsymbol{\omega}$ may be represented by a *mass function* m on Ω, defined as a function from the powerset of Ω, denoted as 2^Ω, to the interval $[0,1]$, such that $m(\emptyset)=0$ and

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

Each number $m(A)$ is interpreted as the probability that the evidence supports exactly the assertion $\boldsymbol{\omega}\in A$ (and no more specific assertion), i.e., the probability of knowing that $\boldsymbol{\omega}\in A$, and nothing more. In particular, $m(\Omega)$ is the probability that the evidence tells us nothing about $\boldsymbol{\omega}$, i.e., it is the probability of knowing nothing. A subset A of Ω such that $m(A)\neq 0$ is called a *focal set* of m. The mass function for which Ω is the only focal set is said to be *vacuous*; it represent total ignorance.

To each normalized mass function m, we may associate belief and plausibility functions from 2^Ω to $[0,1]$ defined as follows,

$$Bel(A)=\sum_{B\subseteq A} m(B) \tag{2a}$$

$$Pl(A)=\sum_{B\cap A\neq\emptyset} m(B), \tag{2b}$$

for all $A\subseteq\Omega$. These two functions are linked by the relation $Pl(A)=1-Bel(\overline{A})$, for all $A\subseteq\Omega$. Each quantity $Bel(A)$ may be interpreted as the probability that the assertion $\boldsymbol{\omega}\in A$ is implied by the evidence, while $Pl(A)$ is the probability that this assertion is not contradicted by the evidence. The function $pl:\Omega\to[0,1]$ that maps each element Ω of Ω to its plausibility $pl(\Omega)=Pl(\{\Omega\})$ is called the *contour function* associated to m.

A key idea in Dempster–Shafer theory is that beliefs are elaborated by aggregating independent items of evidence. Assume that we have two pieces of evidence represented by mass functions m_1 and m_2 on the same frame of discernment Ω. If one piece of evidence tells us that $\boldsymbol{\omega}\in A$ and the other source tells us that $\boldsymbol{\omega}\in B$ for some non-disjoint subsets A and B of Ω, and if both sources are reliable, then we know that $\boldsymbol{\omega}\in A\cap B$. Under the independence assumption, the probabilities $m_1(A)$ and $m_2(B)$ should be multiplied. If, however, A and B are disjoint, we can conclude that the interpretations "$\boldsymbol{\omega}\in A$" and '$\boldsymbol{\omega}\in B$" cannot hold jointly, and the probabilities must be conditioned to eliminate such pairs of interpretations. This line of reasoning leads to the following combination rule, referred to as Dempster's rule [31],

$$(m_1\oplus m_2)(A)=\frac{1}{1-\kappa}\sum_{B\cap C=A} m_1(B)m_2(C) \tag{3a}$$

for all $A\subseteq\Omega, A\neq\emptyset$ and $(m_1\oplus m_2)(\emptyset)=0$, where

$$\kappa=\sum_{B\cap C=\emptyset} m_1(B)m_2(C) \tag{3b}$$

is the *degree of conflict* between m_1 and m_2. If $\kappa=1$, there is a logical contradiction between the two pieces of evidence and they cannot be combined. Dempster's rule is commutative, associative, and it admits the vacuous mass function as neutral element.

Whereas the computation of the full combined mass function $m_1\oplus m_2$ may be prohibitive in very large frames of discernment, the corresponding contour function can be computed in time proportional to the size of the frame, using the following property,

$$pl_1\oplus pl_2=\frac{pl_1 pl_2}{1-\kappa}, \tag{4}$$

where pl_1 and pl_2 are the contour functions of two mass functions m_1 and m_2, and the same symbol $\oplus$ is used for mass functions and contour functions.

2.2. EK-NN rule

Consider a classification problem in which an object o has to be classified in one of c groups, based on its distances to n objects in a dataset. Let $\Omega=\{\omega_1,\ldots,\omega_c\}$ be the set of groups, and d_j the distance between the object to be classified and object o_j in the dataset. If object o_j belongs to group $\omega_{k(j)}$, then the knowledge that object o is at a distance d_j from o_j is a piece of evidence that can be represented by the following mass function on Ω,

$$m_j(\{\omega_{k(j)}\})=\alpha_j, \tag{5a}$$

$$m_j(\Omega)=1-\alpha_j, \tag{5b}$$

with

$$\alpha_j=\varphi(d_j), \tag{5c}$$

where φ is a non-increasing mapping from $[0,+\infty)$ to $[0,1]$, such that

$$\lim_{d\to+\infty}\varphi(d)=0. \tag{6}$$

According to (5), the mass function m_j has two focal sets: the class $\omega_{k(j)}$ of o_j, and Ω. It becomes vacuous when d_j becomes infinitely large. In [6], it was proposed to choose φ as $\varphi(d_j)=\alpha_0\exp(-\gamma_k d_j)$ for some constants α_0 and γ_k. Considering the distances to the n objects in the database as independent pieces of evidence, the n mass function m_j can then be combined by Dempster's rule to yield the combined mass function

$$m=m_1\oplus m_2\oplus\cdots\oplus m_n. \tag{7}$$

For computational reasons, the mass functions m_j for objects o_j that are very dissimilar to object o are nearly vacuous and can be neglected. A useful heuristic is to consider only the *K nearest neighbors* of object o in the database. Denoting by N_K the set of indices of these nearest neighbors, the combined mass function thus becomes

$$m=\bigoplus_{j\in N_K} m_j. \tag{8}$$

If a decision has to be made, one can then assign object o to the class ω_k with the highest plausibility. We can remark that, to make a decision, we need not compute the combined mass function m explicitly. The contour function pl_j corresponding to m_j in (5) is