International Conference on Information Engineering, Management and Security
2015 [ICIEMS 2015]

| | | | |
|---|---|---|---|
| **ISBN** | 978-81-929742-7-9 | **VOL** | 01 |
| **Website** | www.iciems.in | **eMail** | iciems@asdf.res.in |
| **Received** | 10 - July - 2015 | **Accepted** | 31- July - 2015 |
| **Article ID** | ICIEMS022 | **eAID** | ICIEMS.2015.022 |


# A Survey on Pattern classification with missing data Using Dempster Shafer theory

M.Kowsalya[1], Dr.C.Yamini[2]
[1]Research Scholar, Department of Computer Science,
Sri Ramakrishna College of Arts and Science for Women,Coimbatore
[2]Associate Professor, Department of Computer Science,
Sri Ramakrishna College of Arts and Science for Women, Coimbatore

***Abstract:*** *The Dempster-Shafer method is the theoretical basis for creating data classification systems. In this system testing is carried out using three popular (multiple attribute) benchmark datasets that have two, three and four classes. In each case, a subset of the available data is used for training to establish thresholds, limits or likelihoods of class membership for each attribute for each attribute of the test data. Classification of each data item is achieved by combination of these probabilities via Dempster's Rule of Combination. Results for the first two datasets show extremely high classification accuracy that is competitive with other popular methods. The third dataset is non-numerical and difficult to classify, but good results can be achieved provided the system and mass functions are designed carefully and the right attributes are chosen for combination. In all cases the Dempster-Shafer method provides comparable performance to other more popular algorithms, but the overhead of generating accurate mass functions increases the complexity with the addition of new attributes. Overall, the results suggest that the D-S approach provides a suitable framework for the design of classification systems and that automating the mass function design and calculation would increase the viability of the algorithm for complex classification problems.*

***Keywords:*** Dempster-Shafer theory, data classification, Dempster's rule of combination.

## 1. INTRODUCTION

The ability to group complex data into a finite number of classes is important in data mining, and means that more useful decisions can be made based on the available information. For example, within the field of medical diagnosis, it is essential to utilise methods that can accurately differentiate between anomalous and normal data. In DST, evidence can be associated with multiple possible events, e.g., sets of events. The chief aims here are to describe the use of the Dempster-Shafer (D-S) theory as a framework for creating classifier systems, test the systems on three benchmark datasets, and compare the results with those for other techniques.As a result, evidence in DST can be meaningful at a higher level of abstraction without having to resort to assumptions about the events within the evidential set. Where the evidence is sufficient enough to permit the assignment of probabilities to single events, the Dempster-Shafer model collapses to the traditional probabilistic formulation. One of the most important features of Dempster-Shafer theory is that the



**Cite this article as:** M.Kowsalya, Dr.C.Yamini. "A Survey on Pattern classification with missing data Using Dempster Shafer theory." *International Conference on Information Engineering, Management and Security (2015)*: 134-138. Print.

model is designed to cope with varying levels of precision regarding the information and no further assumptions are needed to represent the information. It also allows for the direct representation of uncertainty of system responses where an imprecise input can be characterized by a set or an interval and the resulting output is a set or an interval.

**2.Data classification**

Data classification is the process of organizing data into categories for its most effective and efficient use.
A well-planned data classification system makes essential data easy to find and retrieve. This can be of particular importance for risk management, legal discovery, and compliance. Written procedures and guidelines for data classification should define what categories and criteria the organization will use to classify data and specify the roles and responsibilities of employees within the organization regarding data stewardship Once a data-classification scheme has been created, security standards that specify appropriate handling practices for each category and storage standards that define thedata's lifecyle requirements should be addressed.

**3. Dempster Shafer theory**

The drawbacks of pure probabilistic methods and of the certainty factor model have led us in recent years to consider alternate approaches. Particularly appealing is the mathematical theory of evidence developed by Arthur Dempster. We are convinced it merits careful study and interpretation in the context of expert systems. This theory was first set forth by Dempster in the 1960s and subsequently extended by Glenn Sharer. In 1976, the year after the first description of CF's appeared; Shafer published A Mathematical Theory of Evidence (Shafer, 1976). Its relevance to the issues addressed in the CF model was not immediately recognized, but recently researchers have begun to investigate applications of the theory to expert systems (Barnett, 1981; Friedman, 1981; Garvey et al., 1981). We believe that the advantage of the Dempster-Shafer theory over previous approaches is its ability to model the narrowing of the hypothesis set with the accumulation of evidence, a process that characterizes diagnostic reasoning in medicine and expert reasoning in general. An expert uses evidence that, instead of bearing on a single hypothesis in the original
Equal certainty. Because he attributes belief to subsets, as well as to individual elements of the hypothesis set, we believe that Shafer more accurately reflects the evidence-gathering process. Hypothesis set, often bears on a larger subset of this set. The functions and combining rule of the Dempster-Shafer theory are well suited to represent this type of evidence and its aggregation.

**4. New Method for Classification of Incomplete Patterns**

The new prototype-based credal classification (PCC) method provides multiple possible estimations of missing values according to class prototypes obtained by the training samples. For a c-class problem, it will produce c probable estimations. The object with each estimation is classified using any standard2 classifier. Then, it yields c pieces of classification results, but these results take different weighting factors depending on the distance between the object and the corresponding prototype. So the c classification results should be discounted with different weights, and the discounted results are globally fused for the credal classification of the object. If the c classification results are quite consistent on the decision of class of the object, the fusion result will naturally commit this object to the specific class that is supported by the classification results. However, it can happen that high conflict among the c classification results occurs which indicates that the class of this object is quite imprecise (ambiguous) only based on the known attribute values. In such conflicting case, it becomes very difficult to correctly classify the object in a particular (specific) class, and it becomes more prudent and reasonable to assign the object to a meta-class (partial imprecise class) in order to reduce the misclassification rate. By doing this, PCC is able to reveal the imprecision of the classification due to the missing values which is a nice and useful property. Indeed in some applications, especially those related to defense and security (like in target classification) the robust credal classification results are usually more preferable than the precise classification results subject potentially to a high risk of error. The classification of the uncertain object in meta-class can be eventually precisiated (refined) using some other (costly) techniques or with extra information sources if it is really necessary. So PCC approach prevents us to take erroneous fatal decision by robustifying the specificity of the classification result whenever it is necessary to do it.
A.Determination of c estimations of missing values in incomplete patterns
Let us consider a test data set $X = \{x1, \ldots, xN\}$ to be classified using the training data set $Y = \{y1, \ldots, yH\}$ in the frame of discernment $\Omega = \{\omega1, \ldots, \omega c\}$. Because we focus on 2 In our context, we call standard a classifier working with complete patterns. The classification of the incomplete data (test sample) in this work, one assumes that the test samples are all incomplete data (vector) with single or multiple missing values, and the training data set Y consists of a set of complete patterns. The prototype of each class i.e. $\{o1, \ldots, oc\}$ is calculated using the training data at first, and og corresponds to class ωg. There exist many methods to produce the prototypes. For example, the K-means method can be applied for each class of the training data, and the clustering center is chosen for the prototype. The simple arithmetic average vector of the training data in each class can also be considered as the prototype, and this

**Cite this article as:** M.Kowsalya, Dr.C.Yamini. "A Survey on Pattern classification with missing data Using Dempster Shafer theory." *International Conference on Information Engineering, Management and Security (2015)*: 134-138. Print.

method is adopted here for its simplicity. Mathematically, the prototype is computed for g = 1, . . . , c by og = 1 Tg X yj∈ωg yj (4) where Tg is the number of the training samples in the class ωg.

**5. Basic mathematical terminology and the TBM**

The D-S theory begins by assuming a frame of discernment (Θ), which is a finite set of mutually exclusive propositions and hypotheses (alternatives) about some problem domain. It is the set of all states under consideration. For example, when diagnosing a patient, Θ would be the set consisting of all possible diseases. The power set 2 Θ is the set of all possible sub-sets of Θ including the empty set Φ. For example, if:
Θ = {a, b}
Then
2 = { {, }, { }, Θ}. Θ φ a b
The individual elements of the power set represent propositions in the domain that may be of interest. For example, the proposition "the disease is infectious" gives rise to the set of elements of Θ that are infectious and contains all and only the states in which that proposition is true. The theory of evidence assigns a mass value m between 0 and 1 to each subset of the power set. This can be expressed mathematically as:
2: →[ ].1,0 Θ m
The function (3) is called the mass function (or sometimes the basic probability assignment) whenever it verifies two axioms: First, the mass of the empty set must be zero:
m(φ) = ,0
and second, the masses of the remaining members of the power set must sum to 1:
∑⊆ Θ = A m( A) 1 .
The quantity m(A) is the measure of the probability that is committed exactly to A [3]. In other words, m(A) expresses the proportion of available evidence that supports the claim that the actual state belongs to A but not to any subset of A. Given mass assignments for the power set, the upper and lower bounds of a probability interval can be determined since these are bounded by two measures that can be calculated from the mass, the degree of belief (bel) and the degree of plausibility (pl). The degree of belief function of a proposition A, bel(A), sums the mass values of all the non-empty subsets of A:
bel ( A) m(B).
The degree of plausibility function of A, pl(A), sums the masses of all the sets that intersect A, i.e. it takes into account all the elements related to A (either supported by evidence or unknown):

**6. Advantages and disadvantages of D-S**

The systems described in this paper are all based on the theory presented in Sections 2.1 and 2.2, but D-S-based systems have a great deal of scope and flexibility as regards to system design, which means that classifiers can be created that are highly suited for solving any given problem. In particular, there are no fixed rules regarding how the mass functions should be constructed or how the data combination should be organized For example, consider the case where a car window has been broken and there are three suspects Jon, Mary, and Mike, and two witnesses, W1 and W2. W1 assigns a mass value of 0.9 to "Jon is guilty" and a mass value of 0.1 to "Mary is guilty". However, W2 assigns a mass value of 0.9 to "Mike is guilty" and a mass value of 0.1 to "Mary is guilty". Applying the DRC returns a value of 0.99 for K, which yields a value of 1 for "Mary is guilty". This is clearly counterintuitive since both witnesses assigned very small mass values to this hypothesis. The conflicting beliefs management problem is only a cause for concern when there are more than two classes, so the WBCD dataset used here presents no potential problem. Furthermore, the mass functions used with other two datasets are selected so that any conflicting beliefs are reduced (see Sections 5.2 and 6.2). This is possible since the problem is caused by conflicting mass values, not mass functions, so one can design mass functions and DRC combination strategies that minimize the problem. Some alternative combination rules that attempt to reduce the conflicting beliefs management problem have also been proposed, as in [7] and [8], but none have yet been accepted as a standard method.

**7. Review of D-S applications**

The D-S theory has previously been shown to be a powerful combination tool, but to date most of the research effort has been directed towards using it to unite the results from a number of separate classification techniques. For example, in [30] the results from a Bayesian network classifier and a fuzzy logic-based classifier are combined and in [31] the D-S theory is used in conjunction with a neural network methodology and applied to a fault diagnosis problem in induction motors. The DRC acts as a data fusion tool, i.e. eight faulty conditions are first classified using the neutral network and the classification information is then converted to mass function assignments. These are then combined using DRC, which reduces the diagnostic uncertainty. Al-Ani and Deriche [32] also propose a classifier combination method based on the D-S approach. They propose that the success of the D-S methodology lies in its powerful ability to combine evidence measures from multiple classifiers. In other words, when the results of several classifiers are combined, the effects of their individual limitations as classifiers are significantly reduced. Valente and Hermansky [33] also suggest a DRC methodology that combines the outputs from various neural network classifiers, but in their work it is applied to a multi-stream speech

**Cite this article as:** M.Kowsalya, Dr.C.Yamini. "A Survey on Pattern classification with missing data Using Dempster Shafer theory." *International Conference on Information Engineering, Management and Security (2015)*: 134-138. Print.

recognition problem. As mentioned previously, the work here differs from the above approaches in that it is concerned with classification using the D-S theory alone; no other categorization techniques are employed at any stage in the classification process. This perspective is fairly novel as other works concerned with the D-S theory as a single classifier has mostly focused on adapting its methodology. For example, Parikh et al. [34] present a new method of implementing D-S for condition monitoring and fault diagnosis, using a predictive accuracy rate for the mass functions. The author's claim that this architecture performs better than traditional mass assignment techniques as it avoids the conflicting beliefs assignment problem. In other D-S related work, Chen and Venkataramanan [35] show that Bayesian inference requires much more information than the D-S theory, for example a priori and conditional probabilities. They postulate that the D-S method is tolerant of trusted but inaccurate evidence as long as most of the evidence is accurate.

## 8. The application of D-S to data classification

As discussed in Section 2.3, the D-S theory provides a general framework for creating classifier systems. This framework can be expressed as a series of steps that must be undertaken namely: 1. Define the frame of discernment ($\Theta$). This is the set of all possible hypotheses related to the given dataset and identifies the classes to which the data must be assigned. 2. Determine which data attributes are important for establishing class membership and discard the others. In general, the frame of discernment and the selected attributes (their number and their data types) will provide loose guidelines for designing mass functions and the structure of the DRC combinations. 3. Examine the selected attributes and their data values within a subset of the data in order to design mass functions for each attribute. These functions will be used to assign mass values to the corresponding hypotheses based on the attribute values of the test data. 4. Design a DRC combination strategy based on the data structure. A single application of DRC combines the mass values of each attribute for each data item, but many applications can be used, and DRC can also be used to combine the results of previous applications. 5. Following combination, select a rule that converts the result to a decision. Several may be used on different steps, but the final one ultimately classifies the data.

## 9. Conclusions

This work has utilized the D-S theory (in particular mass functions and DRC) as a framework for creating classification algorithms, and has applied them to three standard benchmark datasets, the WBCD dataset, the Iris dataset, and part of the Duke Outage dataset. For the WBCD, the mass functions were created by considering threshold values in the training data and using a sigmoid model. In this case, classification was a simple one-step process. The accuracy proved to be much higher when all the data attributes were considered (97.6%), and this result was superior to other published results for other popular methods. Furthermore, the D-S method permitted the inclusion of data items that contained missing values in the dataset. Some of the other methods were unable to do this. This paper has hence demonstrated that the D-S approach works well with all three datasets provided the system is designed in the right way and the attributes are carefully selected. Attribute selection appears to influence overall performance considerably, for example, use of all the attributes worked well for the WBCD but not for the Duke Outage data. The D-S theory provides the framework for system design only, and in this sense allows the creation of systems that can be essentially tailored towards the specific problem domain of interest. This may be considered a disadvantage in that there are no strict guidelines for the detailed design of such systems, but it may also be thought of as an advantage, since the flexibility allows for the tweaking and refinement of the system until the desired output levels are reached, especially if this refinement process can be automated in some way. In particular, automating the attribute selection and mass function calculation processes may make the Dempster-Shafer approach an objective and accurate replacement for current state of the art classification systems.

## REFERENCES


[1]. R.J.LittleandD.B.Rubin, StatisticalAnalysisWithMissingData, 2nded.NewYork, NY, USA: Wiley, 2002.
[2]. A.P. Dempster, "Upper and lower probabilities induced by a multivalued mapping", Ann. Math. Statist. 38 (1967) 325-339.
[3]. G. Shafer, "A Mathematical Theory of Evidence", Princeton University Press, Princeton and London, 1976. [4]K. A. Lawrence, "Sensor and Data Fusion: A Tool for Information Assessment and Decision Making", SPIE, Washington, 2004.
[4]. B. Tessem, "Approximations for efficient computation in the theory of evidence", Artificial Intelligence 61 (1993) 315-329.
[5]. K.Jian,H.Chen,andS.Yuan,"Classificationforincompletedatausingclassifierensembles,"inProc.Int.Conf.NeuralNetw.Brain(ICNN &B'05),Beijing,China,Oct.,pp.559–563.
[6]. K.Pelckmans,J.D.Brabanter,J.A.K.Suykens,andB.D.Moor,"Handlingmissingvaluesinsupportvectormachineclassifiers,"NeuralNetw.,vol.18,nos.5–6,pp.684–692,2005.
[7]. P.ChanandO.J.Dunn,"Thetreatmentofmissingvaluesindiscriminantanalysis,"J.Amer.Statist.Assoc.,vol.6,no.338,pp.473–477,1972.
[8]. J.L.Schafer,AnalysisofIncompleteMultivariateData.London,U.K Chapman&Hall,1997.
[9]. O.Troyanskayaetal.,"MissingvalueestimationmethodsforDNAmicroarrays,"Bioinformatics,vol.17,no.6,pp.520–525,2001.
[10]. G.BatistaandM.C.Monard,"AstudyofKnearestneighbourasanimputationmethod,"inProc.2ndInt.Conf.HybridIntell.Syst.,2002,pp.251–260.


**Cite this article as:** M.Kowsalya, Dr.C.Yamini. "A Survey on Pattern classification with missing data Using Dempster Shafer theory." *International Conference on Information Engineering, Management and Security (2015)*: 134-138. Print.

[11]. J.Luengo,J.A.Saez,andF.Herrera,"Missingdataimputationforfuzzyrule-basedclassificationsystems,"SoftComput.,vol.16,no.5,pp.863–881,2012.
[12].[12].D.Li,J.Deogun,W.Spaulding,andB.Shuart,"Towardsmissingdataimputation:Astudyoffuzzyk-meansclusteringmethod,"inProc.4thInt.Conf.RoughSetsCurrentTrendsComput.(RSCTC04),Uppsala,Sweden,Jun.2004,pp.573–579.
[13]. F.FessantandS.Midenet,"Self-organizingmapfordataestimationandcorrectioninsurveys,"NeuralComput.Appl.,vol.10,no.4,pp.300–310,2002.
[14]. Y. Song, J. Huang, D. Zhou, H. Zha and C. Lee Giles, in: J. N. Kok et al. (Eds.), KNN: Informative K-Nearest Neighbor pattern classification, PKDD 2007, LNAI 4702, Springer-Verlag Berlin Heidelberg, 2007, pp. 248–264.

**Cite this article as:** M.Kowsalya, Dr.C.Yamini. "A Survey on Pattern classification with missing data Using Dempster Shafer theory." *International Conference on Information Engineering, Management and Security (2015)*: 134-138. Print.