2015 International Conference on Computational Science and Computational Intelligence

Health Outcome Prediction with Multiple Models and Dempster-Shafer Theory

Michael Bauer
Department of Computer Science
The University of Western Ontario
London, Canada
bauer@csd.uwo.ca

***Abstract*—Can multiple predictive models be combined to predict health care outcomes? In this paper, we explore this question by considering the use of multiple predictive models as "evidence" and formulate a multi-model approach to prediction based on Dempster-Shafer's Theory of Evidence. Given the accuracy measures of multiple models, we propose a formulation of a combined predictive model based on Dempster-Shafer's Theory. We then evaluate this approach on a set of data and compare it to predictions by the individual models.**

***Keywords*—predictive models, model accuracy, Dempster-Shafer Theory.**

I. Introduction

Mathematical models, such as ones based on linear regression, decision trees, etc., are frequently developed in order to try to predict the likelihood or plausibility of an individual having a particular disease or condition given a number of measured parameters about the individual. These parameters can range from discrete values, such "male" or "female", to continuous ones, such as temperature. The broad goal in developing such models is to understand potential factors contributing to a condition and to be able to anticipate or predict potential problems in order to aid in decision making.

The development of such models typically includes an assessment of the accuracy of the model. A portion of the data set used to create the model, or a similar data set, is used to determine the accuracy of the model once it has been developed, i.e., how well it predicts whether the condition is present or not. A perfect predictive model would have an accuracy of 100%, though in most cases, this is not the case.

Suppose, then, that one has a data set and builds two models using different techniques, say model M_1 and model M_2. Suppose further that the accuracy of these models is 90% and 80%, respectively. For a given set of values, suppose that both models predict that the condition is present. This leaves one with a sense of confidence that the condition is present. At least this is more appealing than if one model predicted that the condition was present and the other predicted that it was not. In the latter situation, would one model be preferred over the other, say the one with an accuracy of 90%?

Our interest is in being able to explore the use multiple models for prediction, evaluate the effectiveness of multi-model predictive techniques versus individual models and to have some measure of the overall "likelihood" in the predictions based on multiple models. One approach is to simply rely on joint probabilities. In the previous case, the probability of both models being correct would be 72%; the probability of model M_1 being correct and model M_2 being incorrect would be $0.9 \times (1 - 0.8)$, or 18%, etc. This is certainly one approach, but we would like to have an approach which provides a more intuitive measure of our confidence. Alternatively, one can think of the two models as providing "evidence" for the presence or absence of the condition and we are looking for an estimate of our confidence in that "evidence". We would like to be able to consider the outcome of predictions from two or more different models and have a decision procedure, with some measure of "evidence", i.e., of what the outcome should be based on the results of the individual models.

This same kind of reasoning prompted Arthur Dempster and Glenn Shafer to pursue a rigorous foundation for the notion of evidence. Much of the fundamental ideas appeared in Shafer's seminal work *A Mathematical Theory of Evidence* [1]. Dempster-Shafer Theory (DST), as it is commonly called, is not new to understanding disease, illnesses, etc. It has been used in a variety of scenarios as a means of quantifying evidence associated with diagnosis, both within and away from, the medical area. Our interest here is in the use of the theory with multiple models and, in particular, in using the resulting combined models for prediction.

The remainder of this paper is organized as follows: In Section II, we provide some background on DST and then in Section III we provide an overview of some of the previous work making use of DST in a variety of domains. We introduce our approach to predictive modeling using DST in Section IV and illustrate the approach on an example in Section V. In Section VI we present a brief conclusion and identify some future directions.

II. Basic Concepts of Dempster-Shafer Theory

We begin with an introduction to the basic concepts in Dempster-Safer Theory (DST). Let X be the set of all states under consideration. In our specific case, we will assume that we are dealing with models where they predict that a condition (event, cause, etc.) is either *present (Yes, has the condition)* (Y) or *does not* (N). Let $X = \{Y, N\}$ and let $P(X) = \{\phi, \{Y\}, \{N\}, \{Y, N\}\}$ be the powerset of X. In essence, the result of using a model results in one of two states and possible combinations of the those states.

The DS theory of evidence assigns a *belief mass* (*mass*) to each subset of the powerset: $P(X) \rightarrow [0, 1]$ and is called a *be-*

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.80

lief assignment (or *mass assignment*). The *belief mass* assigns a value to each set representing a certain level of "belief" in the outcomes represented by the set. For example, the set $\{Y, N\}$ could have a non-zero mass assignment representing the belief that the outcome is either "yes" or "no", i.e., undecided. In this way, DST differs from standard probability theory.

A belief assignment must, however, adhere to two basic axioms. A belief assignment m must adhere to the following:

1) The mass of the empty set is 0; $m(\phi) = 0$.
2) The sum of the remaining members of the power set must add up total of 1:

$$\sum_{A \epsilon P(X)} m(A) = 1.$$

The mass $m(A)$ of a given member, A, of the power set expresses the proportion of all relevant and available evidence that supports the possibility that the actual state belongs to A but no particular subset of A, i.e., pertains only to A.

From the mass assignments, the upper and lower bounds of a probability interval can be defined (in the classical sense); these are referred to as *belief* (or *support*) and *plausibility*. These are defined as follows for a set A:

- The belief set $bel(A)$ is the sum of all the masses of subsets of A:

$$bel(A) = \sum_{B|B \subseteq A} m(B).$$

- The plausibility $pl(A)$ is the sum of all the masses of the sets B that intersect A:

$$pl(A) = \sum_{B|B \cap A \neq \phi} m(B).$$

- The probability, $p(A)$, of A is then within the bounds:

$$bel(A) \leq p(A) \leq pl(A).$$

A key component of the DST that we are interested in is how to combine independent sets of mass assignments. In our particular case, this would allow us to combine two models to form a combined set of assignments and, consequently, combined belief and plausibility values, i.e., allows us to estimate the combined value of the evidence from the two models. This, in turn, would allow us to construct a new model which combines the "evidence" from the two models.

This is done using Dempster's *rule of combination*. Assume that we have two models, M_1 and M_2, defined over the same set of states X. There are two belief assignments m_1 and m_2. Then the combination, say a joint model called $M_{1,2}$ and its belief assignments $m_{1,2}$ is defined as follows:

1) $m_{1,2}(\phi) = 0$.
2) The mass of other non-empty sets A is determined by:

$$m_{1,2}(A) = (m_1 \oplus m_2)(A) = \frac{N}{K}, \qquad (1)$$

where, $N = \sum_{B \cap C = A \neq \phi} m_1(B) m_2(C)$,

and $K = 1 - \sum_{B \cap C = \phi} m_1(B) m_2(C)$.

For this computation, B is a set from M_1 and C is a set from M_2. Intuitively, the numerator N computes a measure of evidence from both models which is then normalized by a measure of "conflict" present from other sets, represented by the denominator, K.

III. A Brief Snapshot of Previous Work Utilizing Dempster-Shafer Theory

The notion of relying on "evidence" in decision-making occurs in a variety of domains. In a broad sense, taking evidence (or input) from multiple diverse sources and computing an overall estimate can also be viewed as a means of "data fusing". Not surprisingly, then, Dempster-Shafer Theory (DST) has found applicability in a variety of domains. We look first at several different approaches for using DST and then consider its role in health-related analyses.

Chen [2] looks at using DST to analyze data for intrusion detection. He uses it because it can capture uncertainty and provides a numerical approach for fusing together multiple pieces of evidence from unreliable data, such as data collected from multiple sources or systems while trying to detect intruders. He also notes that it is unlike Bayesian theory in that one does not need to know about a priori or conditional probabilities which are often difficult or impossible to estimate for intruder attacks.

The use of DST to fuse data from multiple sources has been a motivator for a variety of work. Luo [3] considers a variation on DST's computation of evidence with a different algorithm for use in data fusion from multiple sensors. Murphy [4] looks at the utility of DST for sensor fusion for autonomous mobile robots. He looks at DST's weight of conflict metric to measure the amount of consensus between different sensors and evaluated the approach experiments using four types of sensor data collected by a mobile robot. Rakowsky [5] illustrates how to apply DST to safety and reliability modelling.

DST has also been used in information retrieval tasks [6], [7] because there is an opportunity to combine evidence from plausible sources of information dealing with different assessments on the content of documents or articles. Orumchian [6], for example, used DST to fuse information about articles in order to produce an improved qualitative ranking.

Hu et al. [8] looked the use of DST to improve the performance of the use of multiple SVMs: multi-class SVMs (MSVMs) constructed by combining several standard SVM classifiers. The strategy is based on an earlier approach that mapped the output of a SVM to the posterior probabilities [9]. Yen et al [10] showed that the subset-valued mapping used in the Dempster-Shafer theory can be extended to a probabilistic mapping to express the uncertainties and defined a mass function to discount the prior probabilities from each mass. Hu et al build on this work to use the posterior probabilities of multiple SVMs to construct his mass assignments for the SVMs and then combines those using DST.

Given the use of DST within a variety of areas, it is not surprising to find uses within the health/medicine arena, Asland [11] explores the use of DST to combine the classification of three classifiers: k-Nearest Neighbor (kNN), Nave Bayesian and Decision Tree. The combination approach was used to

across two domains: breast tumor classification and skin lesion classification. The belief assignments and uncertainty are designed for each of the individual classifiers. For example, for the Bayesian classifier, posterior probabilities are used to evaluate basic beliefs. DST's rule of combination combines three beliefs to arrive at one final decision. Asland carries out experiments with k-fold cross validation and shows that the nature of the data set has a bigger impact on some classifiers than others and the classification based on combined belief shows better overall accuracy than any individual classifier. He notes that the ability to handle such situations robustly and the ability to classify samples as uncertain in the presence of classifier uncertainty makes this approach attractive for healthcare applications.

Asland's work is similar to the work in this paper - he looks at specific means of combining the results of different classifiers into a single overall classification/decision. Key to his work was specifically designed measures of beliefs and uncertainties, while our work has looked at combining classification evidence from multiple models without specific customization and being able to accomodate many different models.

IV. Predictive Models and DS-Theory

Let us then look at how Dempster-Shafer Theory can help look at the "evidence" provided by multiple different models. For the current work, we will assume that we begin with *accuracy-based models (AB-Models)* defined as $M = (Dom, X, \{p_y, p_n\})$ where Dom is the domain of data for the models (i.e., the actual values used to determine the predictions), $X = \{Y, N\}$ is the set of states (i.e., the model predicts that a condition is present (Y) or not (N)), and p_y and p_n are the measured accuracies of the model in predicting Y and N, respectively. Alternatively, we can think of Y and N as representing two different classes and we have a predictive model that given an instance predicts the class that the instance belongs to. We use Y and N throughout the paper to help provide some intuition on the formation of the models. We begin by looking at how two such AB-Models might be combined to form a joint model based on DST.

A. Forming DS Models

Assume that we have two AB-Models: $M_1 = (Dom, X, \{p_{y1}, p_{n1}\})$ and $M_2 = (Dom, X, \{p_{y2}, p_{n2}\})$. For a $d \epsilon D$, we have $M_j(d) \rightarrow \{Y_j, N_j\}$. That is, each of our models, when given an item, d, from the data space makes a prediction about d - either it satisfies the condition or does not. For any d, there are four possible outcomes:

- M_1 predicts Y and M_2 predicts Y.
- M_1 predicts Y and M_2 predicts N.
- M_1 predicts N and M_2 predicts Y.
- M_1 predicts N and M_2 predicts N.

Consider the first scenario, i.e., for a particular d, $M_1(d) \rightarrow Y$ and $M_2(d) \rightarrow Y$. Based on the predictive likelihood of each of the models, M_1 is likely to predict correctly p_{y1} percent of the time and M_2 is likely to predict correctly p_{y2} percent of the time. But each model could also predict incorrectly with a likelihood of $1 - p_{y1}$ and $1 - p_{y2}$, respectively. We can treat the "evidence" of M_1 predicting correctly as p_{y1}; similarly for M_2. We can also treat $1 - p_{y1}$ as the "evidence" for M_1 being incorrect and similarly for M_2.

These observations allow us to create a DST model for this particular scenario. For the possible outcomes from our two predictive models, we can create a DST-Model corresponding to each, say, $D_1(X, m_1)$ and $D_2(X, m_2)$. We assume that as per the Dempster-Shafer Theory that the possible states associated with each DST model are then $P(X) = \{\phi, \{Y\}, \{N\}, \{Y, N\}\}$ and that for the two models we define the mass assignments as follows:

- $D_1(X, m_1)$
  - $m_1(\phi) = 0$,
  - $m_1(\{Y\}) = p_{1Y}$,
  - $m_1(\{N\}) = 1 - p_{1Y}$,
  - $m_1(\{Y, N\}) = 0$.
- $D_2(X, m_2)$
  - $m_2(\phi) = 0$,
  - $m_2(\{Y\}) = p_{2Y}$,
  - $m_2(\{N\}) = 1 - p_{2Y}$,
  - $m_2(\{Y, N\}) = 0$.

Intuitively, each of these DST models associates a value with the possible result based on the predictive accuracy of the models for **this** scenario, namely, that both models predict *Yes*. Thus, this approach would yield four different models - one for each of the possible outcome scenarios for the two models.

Now that we have both models and associated mass assignments for this scenario, we would like to consider how these might be combined into single model. We would like our combined model to have states that correspond to the outcomes of the predictions of the two individual models. Since we are *assuming* that models predict Y, then we would like the combined model to have some measure (mass) associated with Y in the combined model which reflects the combined evidence.

Let us consider, then, for our current scenario and for the DST models derived from it, how we might combine the models into a single joint model, say $D_{12}(X, m_{12})$. Here, we rely on Dempster's *rule of combination*. The mass of the sets in the combined model is determined by the following:

$$m_{1,2}(A) = (m_1 \oplus m_2)(A) = \frac{N}{K}, \tag{2}$$

where, $N = \sum_{B \cap C = A \neq \phi} m_1(B) m_2(C)$,

and $K = 1 - \sum_{B \cap C = \phi} m_1(B) m_2(C)$.

We consider each of the possible sets in the combined model:

- $m_{12}(\phi) = 0$ (by definition),
- $m_{12}(\{Y\}) = \frac{S_Y}{K_{12}}$

where $S_Y = m_1(\{Y\}) \times m_2(\{Y\}) + m_1(\{Y,N\}) \times m_2(\{Y\}) + m_1(\{Y\}) \times m_2(\{Y,N\})$,

and

$K_{12} = 1 - (m_1(\{Y\}) \times m_2(\{N\}) + m_1(\{N\}) \times m_2(\{Y\}))$,

which is

$= \frac{p_{1Y} \times p_{2Y}}{1-(p_{1Y} \times (1-p_{2Y}) + (1-p_{1Y}) \times p_{2Y})}$

- $m_{12}(\{N\}) = \frac{S_N}{K_{12}}$

  where $S_N = m_1(\{N\}) \times m_2(\{N\}) + m_1(\{Y,N\}) \times m_2(\{N\}) + m_1(\{N\}) \times m_2(\{Y,N\})$

  and K_{12} is as above.

  which is

  $= \frac{(1-p_{1Y}) \times (1-p_{2Y})}{1-(p_{1Y} \times (1-p_{2Y}) + (1-p_{1Y}) \times p_{2Y})}$

- $m_{12}(\{Y,N\}) = \frac{m_1(\{Y,N\}) \times m_2(\{Y,N\})}{1} = 0$

Note that because $m_1(\{Y,N\}) = m_2(\{Y,N\}) = 0$, the intersections of the singletons $\{Y\}$ and $\{N\}$ with $\{Y,N\}$ result in 0 in the numerator and an overall mass of 0 associated with $m_{12}(\{Y,N\})$.

B. Example of Two Models

To illustrate our approach, consider an example in which we have two models which predict whether for some $d \epsilon Dom$ it has a particular condition or not. Our two models are then: $M_1 = (Dom, X = \{Y,N\}, \{0.81, 0.99\})$ and $M_2 = (Dom, X = \{Y,N\}, \{0.73, 0.98\})$. In this case, model M_1 is accurate 81% of the time when predicting that a $d \epsilon Dom$ has the condition and is accurate 99% of the time when predicting that it does not. Similarly, for model M_2 its accuracy is 73% and 98% respectively.

Let us consider Scenario 1, where both predict, for an element d, that d has the condition, i.e., both predict *Yes*. From these models we create a DST-Model corresponding to each, namely, $D_1(X, m_1)_Y$ and $D_2(X, m_2)_Y$; we denote the dependency of the models on the outcome with the subscript Y in both cases. The mass assignments as follows:

- $D_1(X, m_1)_Y$
  - $m_1(\phi) = 0$,
  - $m_1(\{Y\}) = p_{1Y} = 0.81$,
  - $m_1(\{N\}) = 1 - p_{1Y} = 0.19$,
  - $m_1(\{Y,N\}) = 0$.
- $D_2(X, m_2)_Y$
  - $m_2(\phi) = 0$,
  - $m_2(\{Y\}) = p_{2Y} = 0.73$,
  - $m_2(\{N\}) = 1 - p_{2Y} = 0.27$,
  - $m_2(\{Y,N\}) = 0$.

The mass assignments for the combined model, denoted $D_{12}(X, m_12)_{YY}$ where the subscript YY indicates that the model originated from the two models predicting *Yes*, is then:

- $m_{12}(\phi) = 0$,
- $m_{12}(\{Y\}) = \frac{p_{1Y} \times p_{2Y}}{1-(p_{1Y} \times (1-p_{2Y}) + (1-p_{1Y}) \times p_{2Y})}$

  $= \frac{0.81 \times 0.73}{1-(0.81 \times 0.27 + 0.19 \times 0.73)} = 0.9202$
- $m_{12}(\{N\}) = \frac{(1-p_{1Y}) \times (1-p_{2Y})}{1-(p_{1Y} \times (1-p_{2Y}) + (1-p_{1Y}) \times p_{2Y})}$

  $= \frac{0.19 \times 0.27}{1-(0.81 \times 0.27 + 0.19 \times 0.73)} = 0.0798$
- $m_{12}(\{Y,N\}) = 0$

Thus, the combined model suggests that the weight (*belief*, *mass*) of evidence in favor of concluding that d has the condition, namely *Y*, is 0.9202. That is, when both models m_1 and m_2 both predict *Y*, our decision making based on the combined models is to conclude *Y* as well. Intuitively, this makes sense - both models predict *Y*, so why would one conclude anything else?

Let us consider, then, Scenario 2: $M_1(d_i) \rightarrow Y$ and $M_2(d_i) \rightarrow N$. In this case, one model predicts *Y* while the other predicts *N*. What is the decision in the combined model?

We again construct our DST-Models for this Scenario:

- $D_1(X, m_1)_Y$
  - $m_1(\phi) = 0$,
  - $m_1(\{Y\}) = p_{1Y} = 0.81$,
  - $m_1(\{N\}) = 1 - p_{1Y} = 0.19$,
  - $m_1(\{Y,N\}) = 0$.
- $D_2(X, m_2)_N$
  - $m_2(\phi) = 0$,
  - $m_2(\{Y\}) = 1 - p_{2N} = 0.02$,
  - $m_2(\{N\}) = p_{2N} = 0.98$,
  - $m_2(\{Y,N\}) = 0$.

Model m_1 remains the same, but model m_2 changes since the predicted outcome is *N* and the evidence for that has a different value and so the mass assignment changes as well.

We can construct the combined model similarly to the above, but the mass of the combined model, in this case denoted $D_{12}(X, m_{12})_{YN}$, is as follows:

- $m_{12}(\phi) = 0$,
- $m_{12}(\{Y\}) = \frac{p_{1Y} \times (1-p_{2N})}{1-(p_{1Y} \times p_{2N} + (1-p_{1Y}) \times (1-p_{2N}))}$

  $= \frac{0.81 \times 0.02}{1-(0.81 \times 0.98 + 0.19 \times 0.02)} = 0.0800$
- $m_{12}(\{N\}) = \frac{(1-p_{1Y}) \times p_{2N}}{1-(p_{1Y} \times p_{2N} + (1-p_{1Y}) \times (1-p_{2N}))}$

$= \frac{0.19 \times 0.98}{1-(0.81 \times 0.98 + 0.19 \times 0.02)} = 0.9200$

- $m_{12}(\{Y, N\}) = 0$

In this case, the "evidence" in the combined model is for a decision of *N*. This is still intuitive in that the weight of the accuracy of model m_2 when predicting a *N* is so high (0.98) that the "logical" choice is to simply follow model m_2.

The models for the remaining two Scenarios, namely that M_1 predicts N and M_2 predicts Y and that both predict N, follow similarly.

Note also that this approach can be extended to accommodate any number of *AB-Models*, though the number of corresponding *DST-Models* grows exponentially; i.e., for 3 *AB-Models* there are 8 *DST-Models*, for 4 *AB-Models* there are 16 *DST-Models*, etc. Given 3 *AB-Models*, the *DST-Models* would be YYY (all predict Y), YYN (the first two models predict Y and the third predicts N), etc. Fortunately, for even a modest number of models, the mass assignments of all the models can be computed fairly efficiently using an iterative algorithm.

C. Predicting with DS Models

How does one use the DS-Models once they have been constructed? Assume that we have $n\ AB-Models = \{M_1, ..., M_n\}$ over some domain Dom. Assume that we have constructed our $DST-Models = \{DS_1, ..., DS_N\}$ where $N = 2^n$. For predicting the outcome associated with $d \epsilon Dom$, we introduce the following algorithm:

Dominant-Belief Algorithm (DB):
Input: d, *AB-Models* $= \{M_1, ..., M_n\}$, *DST-Models* = $\{DS_1, ..., DS_N\}$

1) Evaluate each of $M_1, ..., M_n$ on d, $M_i(d) = O_i \epsilon (Y, N)$ for each of the models.
2) Find the *DST-Model*, say DS_i corresponding to the predicted outcomes, i.e., the *DST-Model* corresponding to $O_1, ...O_n$.
3) Choose the outcome based on the highest belief value of the outcomes in DS_i.

DB is a simple algorithm to predict a single overall outcome based on the outcomes from the individual models and using their results to select a *DST-Model* which is then used to choose an outcome; we illustrate this in the next section. More importantly, the approach provides a means of quantifying choices based on the outcomes of the original predictive models. This is certainly not the only way to interpret the results. Recall that the "mass" in a *DST-Model* represents "belief", so that the more mass the stronger "belief". This can also be used as part of the decision making; we will also illustrate this in the next section.

V. Application of the Models

In the following we report on the use of the models using a data set dealing with the classification of the severity of mammographic mass lesions as benign or malignant[12]. The data set contains a patient's age, four attributes from mammographic images and the resulting outcome (benign or malignant). There are 961 instances with 516 being benign and 445 being malignant.

	Logistic	*SVM*	*DTree*
Benign Accuracy	0.700	0.685	0.681
Malignant Accuracy	0.840	0.845	0.833

TABLE I. Prediction Accuracy of Logistic, SVM and Decision Tree Models

The data was analyzed using three classification schemes: Logistic Regression (LR), Support Vector Machine (SVM) and Decision Tree (DT); the Weka analysis package [13] was used for the analyses. The models were built using 70% of the data set as training data and the remaining 30% for testing and measuring the predictive accuracy of each method. The results of the prediction accuracy for the methods are summarized in Table I.

The results of computing the *DST-Models* for just LR and SVM are summarized in Table II and for the three models, the results of the *DST-Models* are summarized in Table III. In the table Y is *Benign* and N is *Malignant*.

For the LR and SVM models, the corresponding four *DST-Models* and the masses (belief) associated with outcomes Y and N are provided in Table II. As per the algorithm described in the previous section, given an instance from the domain, a prediction from the LR and SVM models is obtained, either Y or N, and then the corresponding entry in Table II is examined. The "prediction" is then based on the class with the highest mass. For example, if for an instance d, LR predicts Y and SVM predicts N, then using the *DB* algorithm described, we would use the *DST-Model* in the second row of Table I and choose N as our prediction as its mass is 0.700.

Using this approach we evaluated the *DB* algorithm on our test set. The overall prediction accuracies for LR and SVM were respectively 77.8% and 76.4% and using the *DB* algorithm with the DST models the result was 78.9%. The prediction performance is somewhat better than the individual models.

We did a similar analysis using three models, adding a Decision Tree model to our previous two. This resulted in eight *DST-Models* and associated masses; see Table III. Intuitively, one might take a naive approach and just assume that when any two of the three models agree on a choice, then that should be the overall choice. However, computing the masses for the different models using the rule of combination, the results are a little different in this case. Using our very simple algorithm and just relying on the magnitude of the masses, we see that from Table III only two choices would yield a Y choice, namely when all three of the models predict Y or when both LR and SVM predict Y. In all other cases, the prediction is N. Evaluating the accuracies on our test set, we have LR and SVM as before and DT with an accuracy of 75.7%; prediction with the DST models and our simple algorithm achieved an accuracy of 78.9%. In this case the result of having the three models is not any better than having just the two; the Decision Tree model did not add much additional *evidence* in the additional DST models. If one did take the naive approach of just using a voting approach (i.e., choose the response that at least two of the three models agreed upon),

LR	*SVM*	*DST Model*
Y	*Y*	Y = 0.835, N = 0.165
Y	*N*	Y = 0.300, N = 0.700
N	*Y*	Y = 0.293, N = 0.707
N	*N*	Y = 0.034, N = 0.966

TABLE II. DST MODELS BASED ON LR AND SVM MODELS

LR	*SVM*	*DTree*	*DST Model*
Y	*Y*	*Y*	Y = 0.916, N = 0.085
Y	*Y*	*N*	Y = 0.504, N = 0.496
Y	*N*	*Y*	Y = 0.478, N = 0.523
Y	*N*	*N*	Y = 0.079, N = 0.921
N	*Y*	*Y*	Y = 0.467, N = 0.531
N	*Y*	*N*	Y = 0.077, N = 0.923
N	*N*	*Y*	Y = 0.069, N = 0.931
N	*N*	*N*	Y = 0.007, N = 0.993

TABLE III. DST MODELS BASED ON LR, SVM AND DT MODELS

then the result would have been 77.4% – better than the SVM and DT models, but somewhat poorer than that LR model and poorer than the *DB* algorithm.

As noted, the "mass" represents a measure of "belief" so that one does not necessarily have to treat a *DST-model* as purely predictive. In looking at the mass values in Table II the largest magnitudes range from 0.700 to 0.966. One might choose to rely only on "beliefs" at a certain level. We considered a somewhat alternative algorithm: we only consider predictions using the DST model in Table II when LR and SVM both predict Y or both predict N; one can think of instances where this is not the case as "undecided". This resulted in 30 instances not being classified (i.e., "undecided") and the remaining being predicted with an accuracy of 79.8%.

A similar "prediction" algorithm was used with the DST-Model in Table III. Here, we only considered combinations of the LR, SVM and DTree models where there was a mass greater than 0.8. In this case, there were 6 instances (of 288) that were not classified; the remaining 282 were classified with an accuracy of 78.7%.

VI. CONCLUSION AND DIRECTIONS

This approach to the use of Dempster-Shafer Theory provides a means of considering multiple predictive models as "evidence" and can yield alternative approaches for decision making; the results of experiments suggest that there is potential value. The approach presented in this paper, however, is only an initial look at how Dempster-Shafer Theory might be used; a number of further areas are open for study.

The approach taken for combining the results of different classifiers is not dependent on particular models and can easily be extended to include the results of a large number of models. As well, the methodology used to form *DST-Models* is not necessarily the only way to do so from accuracy-based models; other alternatives could be considered. Though the focus here was on predictions over two classes (Y and N), there is no inherent limitation on considering predictions over some finite number of classes. This would broaden the scope of the *DST-Models*.

A particular advantage of Dempster-Shafer Theory is that it allows for the modeling of "uncertainties"; i.e., we could consider predictive models which predict "yes", "no", "unknown", for example, as Asland has done. This would require some changes in the construction of predictive models, but would allow for a formulation of *DST-Models* that could associate a measure of "belief" around these choices. This could help support decision making processes by explicitly capturing confidence around yes-no predictions and confidence that a prediction of "maybe" might suggest further information is needed. We have not yet explored how we might capture uncertainty, as did Asland, though this is something to consider.

Finally, further evaluation of the approach with more and different data, and more experiments in different settings is required.

ACKNOWLEDGMENT

Support for this work was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC)

REFERENCES

[1] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press, 1976.

[2] T. Chen and V. Varadharajan, "Dempster-shafer theory for intrusion detection in ad hoc networks," *IEEE Internet Computing*, vol. 9, no. 6, pp. 35–41, 2005.

[3] Z. Luo and D. Li, "Multi-source informaiton integration in intelligent systems using the plausibility measure," in *IEEE Int. Con. Multisensor Fusion and Integration for Intelligent Systems*. IEEE, 1994, pp. 403–409.

[4] R. R. Murphy, "Dempster-shafer theory for sensor fusion in autonomous mobile robots," *IEEE. Trans. Robotics and Automation*, vol. 14, no. 2, pp. 197–206, 1998.

[5] U. K. Rakowsky, "Fundamentals of the dempster-shafer theory and its applications to reliability modelling," *Int. J. Reliability, Quality and Safety Engineering*, vol. 14, pp. 579–601, 2007.

[6] F. Orumchian, B. N. Araabi, and E. Ashoori, "Using plausible inferences and dempster-shafer theory of evidence for adaptive information filtering," in *4th International Conference on Recent Advances in Soft Computing*, 2002.

[7] M. Lalmas and M. Ekaterini, "A dempster-shafer indexing for focussed retrieval of a hierarchically structured document space: Implememtation and experiments on a web museum collection," in *6th RIAO Conference on Content-Based Multimedia Information Access*, 2000.

[8] Z. Hu, Y. Li, Y. Cai, and X. Xu, "Method of combining multi-class svms using depmpster-shafer theory and its application," in *Proceedings of the Ammerican Control Conference*. AACC, 2005, pp. 1946–1950.

[9] A. Smola, P. Bartlett, B. Scholkopf, and D. Schuur-mans, *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. MIT Press, 1999.

[10] J. Yen, "Dempster-shafer theory for intrusion detection in ad hoc networks," *IEEE Internet Computing*, vol. 9, no. 6, pp. 573–585, 2005.

[11] Y. Aslandogan and G. Mahajani, "Evidence combination in medical data mining," in *Proceedings of Int. Conf. on Information Technology: Coding and Computing*. IEEE, 2004.

[12] R. Schulz-Wendtland, "Mammographic mass data," Machine Learning Repository, University of California, Irvine, Institute of Radiology, Gynaecological Radiology, University Erlangen-Nuremberg, Universittsstrae 21-23, 91054 Erlangen, Germany, Oct. 2007.

[13] T. U. of Waikato Machine Learning Group. (2014) Weka 3: Data mining software in Java. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/index.html