Combination of Type III Digit Recognizers using the Dempster-Shafer Theory of Evidence

Catalin I. Tomai and Sargur N. Srihari
Center of Excellence for Document Analysis and Recognition (CEDAR)
Department of Computer Science and Engineering
520 Lee Entrance, Suite 202, Amherst, NY 14228-2567
{catalin,srihari}@cedar.buffalo.edu

Abstract

We investigate the combination of Type-III classifiers using the Dempster-Shafer Theory of Evidence. Various methods of building BPA's for each classifier using both "global" and "local" classifier information are explored. We propose modifications to two established BPA-computation methods to make them better suited for combining Type-III classifiers. We also show the effectiveness of using compound hypotheses when a classifier cannot confidently choose between the top two returned classes. Experimental tests demonstrate the superiority of some of the approaches proposed here on the numeral recognition problem when combining three different character recognizers with Type-III classification engines.

1 Introduction and Previous Work

Classifier combination has become a common approach to improve classification performance. Various combination methods have been proposed ([5]). Classifiers can be categorized based on their output information levels ([8]) into three types: (i) those that return a unique class label indicating the most probable class to which the input pattern belongs (Type I) (ii) those that return a complete or partial ranked list class labels (syntactic classifiers - Type II) (iii) those that return output information at measurement level (Type III). While Type III classifiers provide more information, because they are typically built using different learning algorithms, their combination faces the problem of combining outputs in different numbers (e.g. a classifier returning only the top three choices) and on different scales.

The Dempster-Shafer Theory of Evidence ([1],[2]) is a proven method for combining information from different sources, whose performance, however, depends very much on the methods used to compute the Basic Probability Assignment (BPA) functions (masses).

In the DS Theory context we need to harmonize the classifier output in a common belief framework, using different conversion functions, depending on the classifiers particularities. Some of the methods proposed in the literature ([8]) avoid transforming the output scores into beliefs by using "global" (a priori) information given by the recognition, substitution and rejection rates of the classifiers. They present the disadvantage of not using all the output information available. Others ([7]) use "local" information, given by the confidence values (output scores).

A classifier combination scheme that takes into consideration the different behavior of the participating classifiers and uses both "local" and "global" information is proposed. Also, modifications to existing methods are proposed to make them more suitable for handling combination of Type-III classifiers.

In many cases the true class can be found in the top two or three choices returned by the classifier, which means that the classifier is highly confident that the true class is the first or second choice, but **uncertain** regarding it's correct identity. Since the DS Theory provides a flexible mechanism to include such uncertainty in the combination process we also investigate the usage of non-atomic hypotheses as focal elements.

To summarize, this work has several objectives:

- Propose a set of methods that combine "global" and "local" information to build classifiers' BPA.
- Propose modifications to classic methods for computing the BPA of a classifier suited for combining Type III classifiers and evaluate their performance.
- Investigate the effect of the use of double hypotheses on classification performance

The paper is organized as follows: Section 2 presents the basic terminology used by the Dempster-Shafer Theory

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

of evidence as well as the parameters used to describe the classifier's performance, as applicable to this theory. Section 3 presents and proposes different methods of computing the BPA's. Section 4 presents some experiments using these methods on the digit and character recognition problem. Section 5 presents the conclusions and some ideas for future work.

2 Terminology

The Dempster-Shafer Theory of Evidence The Dempster-Shafer Theory of Evidence(DST) is a generalization of the Bayesian reasoning used to represent and combine evidences. We present its basic concepts using the terminology from [8]. Let Θ be a set of M exhaustive and mutually exclusive propositions(hypotheses) A_i. $\Theta = \{A_1, A_2, ...A_M\}$. Θ is called *frame of discernment.* All possible subsets of Θ form a superset 2^Θ, each subset $A \subset \Theta$ being an element of 2^Θ. A function m is called a *basic probability assignment*(BPA) if

$$m : 2^\Theta \to [0,1], m(\phi) = 0 \quad and \quad \sum_{A \subseteq \Theta} m(A) = 1 \quad (1)$$

A BPA represents the impact of each evidence on the subsets of Θ.

A *belief function Bel* corresponding to a specific BPA m represents the total belief committed to a subset of Θ.

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

When two or more evidences exist, the Dempster rule of combination can be used to combine them into a new BPA and $Bel(.)$.

Let m_1, m_2 and m be the BPA's for Bel_1,Bel_2 and Bel. The Dempster rule computes m which represents the combined effect of the two evidences m_1 and m_2 as follows:

$$m(A) = m_1 \oplus m_2(A) = K \sum_{X \cap Y = A} m_1(X) m_2(Y)$$

$$K^{-1} = 1 - \sum_{X \cap Y = \phi} m_1(X) m_2(Y) = \sum_{X \cap Y \neq \phi} m_1(X) m_2(Y)$$

where $X \subseteq \Theta, Y \subseteq \Theta$

Type-III Classifiers For K classifiers, $e_1, e_2, ...e_K$, their measured performance on the training set is given by: $\epsilon_r^{(k)}$ (recognition rate) and $\epsilon_s^{(k)}$ (substitution rate). Usually $\epsilon_r^{(k)} + \epsilon_s^{(k)} < 1$ due to the rejection of some samples.

Given an input vector $\bar{x}$, each classifier e_k returns as result an output vector $\bar{y}_k \in R^M$, $\bar{y}_k = e_k(\bar{x})$, where M represents the number of classes (e.g. for digits M = 10). The top class returned by e_k is denoted by e_k^t and the true class is T.

Given the set of M possible classes $C = \{C_1, C_2, ..., C_M\}$, we consider the following *frame of discernment*: $\Theta = \{A_1, A_2, ..., A_M\}$, where $A_i = \bar{x} \in C_i$ which denotes that the input vector $\bar{x}$ belongs to class C_i.

For each classifier e_k we compute the following values from the classifiers outputs for the n_i samples belonging to class i from the training set:

$min_k^i = min(y_k^i) \quad max_k^i = max(y_k^i)$ - the minimum and maximum confidence values

$d_k^i = y_k^i - min_k^i$ - the distance between the confidence value and the minimum confidence value

$dmin_k^i, dmax_k^i$ - the minimum and maximum distance value.

ϕ_k - a mapping (membership function) that maps the confidence value y_k^i to a belief value $\phi_k(y_k^i)$ in the [0-1] interval.

3 Computation of Evidences

Several methods to compute the evidences were proposed in the literature. We can group them based on the level the information used to compute the evidences is extracted.

3.1 Classifier and Class level

Classifier level Xu et al. ([8]) has used the $\epsilon_r^{(k)}$ and $\epsilon_s^{(k)}$ rates estimated on a training set to build the BPA's. Parikh et al([6]) used the predictive rate to compute the beliefs. We refined this method by using the recognition and substitution rate for each class. In their approach the BPA m_k for classifier e_k which returns the top class e_k^t contains two focal elements $A_{e_k^t}$ and $\neg A_{e_k^t} = \Theta - \{A_{e_k^t}\}$, with

$$m_{s(k)}(A_{e_k^t}) = \epsilon_r^{(k)}$$
$$m_{s(k)}(\neg A_{e_k^t}) = \epsilon_s^{(k)}$$
$$m_k(A_{e_k^t}) + m_k(\neg A_{e_k^t}) + m_k(\Theta) = 1$$

We are using the "global" performance values of each classifier (the apriori performance) as a separate source ($s(k)$) of information in the combination mix with the corresponding mass: $m_{s(k)} = m_k$.

Class level Rogova ([7]) proposes a method that computes a mean vector $\bar{E}_i^k$ for each classifier e_k and class C_i from the training set, which is used as a reference vector for that class. A distance $D_i^k = \Phi(\bar{E}_i^k, \bar{y}^k)$ is computed between the output vector and the reference vector and interpreted as evidence pro-hypothesis A_i. In this method the evidence $v_i \bar{y}^k$ for class C_i and classifier e_k is obtained from combining simple support functions with focus A and $\neg A$.

Unlike the previous method, this one requires the conversion of the output values into belief values in the [0,1] range. Also, for every class C_i (not only for the e_k^t) we have

$$\begin{aligned}&m_i(A_i) = D_n^i\\&m_i(\Theta) = 1 - D_n^i\\&m_i(A_i) + m_i(\Theta) = 1\\&m_{\neg i}(\neg A_i) = 1 - \prod_{l \neq i}(1 - D_l^k)\\&m_{\neg i}(\Theta) = 1 - m_{\neg i}(\neg A_i)\\&m_{\neg i}(\neg A_i) + m_{\neg i}(\Theta) = 1\end{aligned}$$

Classifier and Class level The methods that view the classifier's performance only in terms of recognition and substitution rates (classifier level) may miss some important information. Imagine that a classifier e has a recognition rate r and that for two different input vectors x and y the output confidences for the top class are p_x and p_y, with $p_x >> p_y$. Intuitively, there is more belief in the correctness of the classification of x than of y, however, this is not reflected in the computation of their corresponding BPA's.

Reversely, consider classifiers e_1 and e_2 that present recognition rates of r_1 and r_2, with $r_1 > r_2$. Let's assume that for two different input vectors x and y the output confidences for the top class k are the same $p_x = p_y = \alpha$, with the confidences for the other classes equal as well. Using methods that only consider the confidences output (missing the global perspective) the evidences for class k are the same for each classifier, even if we should be more confident in c_1 given its past performance.

We expect that combination methods that take into consideration both types of information to perform better than those that consider only one type of information. We'll verify this assumption experimentally in Section 4. Considering the observations above we have experimented with two different ways of computing the BPA's:

Method 1 For this method, for each classifier e_k, the sum of masses **for all classes** add up to 1 :

$$\begin{aligned}&m_k(A_i) = \phi_k(y_k^i)\\&m_k(\Theta) = 1 - \sum_{A_i \subset \Theta} m_k(A_i)\end{aligned}$$

Here masses are computed for each class and derived from the output confidences,similar to [7] and unlike in [8] where masses are computed only for the top class and derived from 'global information. We may add the "global" level information by introducing the $s(k)$ sources in the combination mix.

Method 2 In [7], unlike the previous method, the sum of masses for **one class** add up to 1. Combining the knowledge about A_i we obtain evidence v_i:

$$\begin{aligned}&v_i(\bar{y}^k) = m_i \oplus m_{\neg i}(A_i) = \frac{D_i^k \prod_{l \neq i}(1-D_l^k)}{1-D_i^k[1-\prod_{l \neq i}(1-D_l^k)]}\\&v_i^k(\Theta) = (1 - D_i^k)\prod_{l \neq i}(1 - D_l^k)\end{aligned}$$

In the final step, a measure of confidence for each class C_i is computed as follows: $v_i(\bar{x}) = v_i(y^1) \oplus ... \oplus v_i(y^K)$, which, after a normalization step becomes: $v_i(\bar{x}) = C\prod_n v_i(y^n)$, where C is a constant.

The larger the value of M the more training data we need to obtain a meaningful mean vector for each class. Therefore, we decided to use the confidence values returned by the classifiers as beliefs in case they are in the [0-1] interval eliminating the possibly less accurate conversion efforts. If not, we use a membership function that returns evidence values based on the output confidence value y_k^i, not on the entire vector $\bar{y}_k$ like in the method described above. Because of these adjustments, in the final step, instead of multiplying the $v_k(y^n)$ values (after normalization), we follow the traditional DST combination scheme and use $v(\Theta)$ when computing the sum $v_i(y^m) \oplus v_i(y^n) = v_i(y^m)v_i^n(\Theta) + v_i(y^n)v_i^m(\Theta) + v_i(y^m)v_i(y^n)$. Since we don't use the proximity measures anymore, we may have cases in which $v_i(y^m) = 0$, which if multiplied with the rest of evidences for class C_i in the original scheme reduces all belief to 0, an undesirable outcome.

3.2 Compound hypotheses

For Type III classifiers, measurement values are usually returned not only for the top class but also for the rest of the classes. However, most of the times the right class can be found among the top two or three choices returned by the classifier.

We can interpret this behavior as uncertainty regarding the true identity of the true class and confidence that the true class is found among this small number of candidates. Uncertainty about the right class cannot be easily modeled using the Bayesian reasoning, where a certain probability has to be assigned to each individual class. DST allows us to assign measures of support to composite hypotheses (e.g $C_i \vee C_j$), to express the uncertainty regarding which one of these classes is the right one.

For our test case this behavior can appear only for classifier e_1 since e_2 and e_3 always assign a high confidence to the top class and a low confidence to the other classes.

For example, for classifier e_1 we can have the following case: The confidences in the top two classes are o_i = 4.5, o_j = 5.7. Considering that the "best"(lowest) confidence values for the two classes over the training set are min_1^i = 4.3 and min_1^j = 5.3, we can say that the classifier is simultaneously highly confident in both classes.

If we use compound (double) hypotheses the BPA computation changes accordingly:

$$m_k = \begin{cases} \sum_{A_l \subset \Theta} m(A_l) + m(A_i \cup A_j) + m(\Theta), \\ \qquad\qquad if \quad |\bar{y}_k^i - \bar{y}_k^j| < \mu_k \quad and \quad C_i = e_k^t \\ \sum_{A_l \subset \Theta} m(A_l) + m(\Theta), \qquad else \end{cases}$$

where the threshold values μ_k are determined from the training set.

While multiple hypotheses are usually avoided for considerations of efficiency, several methods to compute the corresponding BPA were already proposed. When two classes C_i and C_j are not distinguishable, according to [3] two strategies can be chosen:

- $m(A_i) = m(A_j) = 0$ and $m(A_i \cup A_j) \neq 0$
- $m(A_i) = m(A_j) = m(A_i \cup A_j) \neq 0$

A different approach is used in [9] where the the mass assigned to double hypothesis $C_i \cup C_j$ is proportional to the fuzzy membership of the sample in question to both classes C_i and C_j. The mass value represents the surface of a triangle that depends on the degrees of the membership of the sample (e.g. pixel) to the two classes and on the difference between these two degrees.

In our case we do not have membership functions, but scores that indicate a lower or higher confidence in the sample's membership in a certain class. After experimenting with several methods of dividing the confidence between C_i, C_j, $C_i \cup C_j$ and the other classes, we have equally divided the evidence among $m(A_i)$, $m(A_j)$, $m(A_i \cup A_j)$ and $m(\Theta)$. This distributes the uncertainty equally among the possible choices.

4 Results and Analysis

Recognizers Description The recognizers considered are not homogeneous in that their feature extraction and classification engines are different: (i) C_1 - Returns ranked confidences(distances) for all N classes. Values are on a specific scale. (ii) C_2 - Returns only the top two classes with attached confidences. Values add up to 1. (iii) C_3 - Returns ranked confidences for all N classes, however, the confidence in the top class is usually extremely high compared with the confidences for the rest of the classes.

Mapping functions Because of the lack of homogeneity of the recognizers output, to map the output values into evidence values in the [0-1] interval we have to use different mapping functions, depending on the characteristics of each classifier. In our case, two of the classifiers already return values in the [0-1] interval and can be therefore used directly as beliefs. The effect of smoothing the output of the third recognizer is currently under investigation.

$$\phi_k(y_k^i) = \begin{cases} 1 - \frac{d_k^i - dmin_k^i}{dmax_k^i - dmin_k^i} & for \quad k = 1 \\ y_k^i & for \quad k = 2,3 \end{cases}$$

In the end the evidences for each class C_i for each recognizer are combined according to the Dempster rule described in Section 2. If we consider all sources of information described above we will have:

$$m = m_k \oplus m_{s(.)}$$

where m_k may or may not include evidences for double hypotheses.

The input vector $\bar{x}$ is assigned to class C_j if $m(A_j) = max_{1 \leq l \leq M} m(A_l)$. While the recognizers considered here do not have a reject mechanism ($\epsilon_r^{(k)} + \epsilon_s^{(k)} = 1$), in the DST framework rejection appears when sources of information conflict (The value of K is 0), therefore providing an automatic rejection mechanism for the combination scheme.

Results The experiments have been performed on several sets of character images obtained from two standard sets (CEDAR and NIST) and a CEDAR internal set composed of digits collected from handwritten postal addresses. We report here the results obtained on one of the sets (results obtained on the other sets were similar) which was divided into a training and testing subsets of size 25656 and respectively 12242 images. Table 1 presents the results obtained using the following methods.
X_1 - original method presented in [8]
X_2 - unlike method X_1 we use the substitution and recognition rate **for each class** in computing the BPA's
X_3 - unlike method X_1 the predictive rate is used to compute the BPA's (see [6]).
R_1 - original method presented in [7] using the cosine measure as the proximity measure.
R_2 - original method presented in [7] using a function based on the Euclidean distance as the proximity measure
BKS - the "behavior-knowledge space" method ([4]).
M_1,M_2 - described in Section 3

Analysis Since the performance of recognizer e_3 is already high, the improvements are bound to be marginal. However, from the results presented in Table 1 we can draw some preliminary observations.

- Using double hypotheses indeed helps improving the classification performance.
- The "global" information hurts more than helps the performance, especially for M_2. This is due to the

Method	Recogn	Error	Reject
C_1	0.948	0.052	0.000
C_2	0.840	0.160	0.000
C_3	0.977	0.023	0.000
X_1	0.979	0.021	0.000
X_2	0.965	0.035	0.000
X_3	0.980	0.020	0.000
R_1	0.975	0.025	0.000
R_2	0.974	0.026	0.000
BKS	0.979	0.015	0.006
M_1	0.980	0.020	0.000
M_1 + s_k	0.980	0.020	0.000
M_1 + dh	0.982	0.017	0.001
M_1 + dh +s_k	0.971	0.015	0.014
M_2	0.979	0.021	0.000
M_2 + s_k	0.845	0.155	0.000

Table 1. Recognition rates for different combination methods and different reject rates

D	**0**	1	2	9	1∪2	Θ
C_1	9.1811	7.7430	12.25	8.824	-	-
C_2	0	0.4893	0	0	-	-
C_3	0.8263	0.0008	0	0.001	-	-
$m_1(X_1)$	0	0.9512	0	0	-	0.0488
$m_1(M_1+dh)$	0	0.25	0	0.25	0.25	0.25
$m_2(X_1)$	0	0.8560	0	0	-	0.1440
$m_2(M_1+dh)$	0	0.4893	0	0	-	0.51
$m_3(X_1)$	0.9878	0	0	0	-	0.0122
$m_3(M_1+dh)$	0.8263	0.0008	0	0.001	-	0
$m_c(X_1)$	0.3774	**0.6180**	0	0	-	0
$m_c(M_1+dh)$	**0.440**	0.308	0	0.08	0	0.0789

Table 2. Example of applying the methods X_1 and M_1 with double hypotheses for combining the outputs of the C_1, C_2 and C_3 recognizers

normalization step, which, given that we have mass assigned only for the top class and Θ, builds BPA's that have $m(A_i) = 0, \forall i \neq e_k^t$ and $m(A_{e_k^t}) = 1$. The combination of this evidence with the evidences from the other sources introduces a bias towards one single class that cancels the evidences for the other classes.

- The M_1 set of methods present more stable performances as compared with the M_2 set. They also present a better performance than the established X_1,R_1,R_2 methods.

In Table 2 we can see how two methods(X_1 and M_1 with double hypotheses) are applied on one example. The first three rows contain the recognizer's output values for the top four digits (0, 1, 2 and 9). The remaining rows contain the values of the masses for each classifier and each method (m_1,m_2,m_3) and their combination (m_c). Method X_1 returns the wrong class(1) while M_1, which by using double hypotheses distributes the belief obtained from C_1 between the top two classes, their reunion and the Θ, return the true class (numeral 0).

5 Conclusions and Future Work

We have explored improvements to classical methods for building BPA's using "global" and "local" classifier information together. We have experimented with the use of double hypotheses in cases in which a classifier cannot choose between two classes. The proposed method compared favorably with other DST-based methods proposed in the literature for the problem of digit classification when combining three character recognizers with Type-III classification engines on large sets of data.

Future work would explore different ways of making the final decision based on the combination BPA's and include as further experimentation on letter images.

References

[1] AP.Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

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

[3] S. L. Hegarat-Mascle, I. Bloch, and D. V. Madjar. Application of dempster-shafer evidence theory to unsupervised classification in multisource remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 35(4):1018–1031, 1997.

[4] Y. S. Huang and C. Y. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):90–94, 1995.

[5] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

[6] C. R. Parikh, M. J. Pont, and N. B. Jones. Application of dempster-shafer theory in condition monitoring systems: A case study. *Pattern Recognition Letters*, 22(6-7):777–785, 2001.

[7] G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7(5):777–781, 1994.

[8] L. Xu, A. Krzyzak, and C. Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3), 1992.

[9] Y. M. Zhu, L. Bentabet, O. Dupuis, V. Kaftandjian, D. babot, and M. Rombaut. Automatic determination of mass functions in dempster-shafer theory using fuzzy c-means and spatial neighborhood information for image segmentation. *Optical Engineering*, 41(4):760–770, 2002.