Evidence Combination in Medical Data Mining

Y. Alp Aslandogan, Gauri A. Mahajani
Department of Computer Science and Engineering, The University of Texas at Arlington
alp@cse.uta.edu
Stan Taylor
Department of Dermatology, The University of Texas Southwestern MedicalCenter

Abstract

In this work we apply Dempster-Shafer's theory of evidence combination for mining medical data. We consider the classification task in two domains: Breast tumors and skin lesions. Classifier outputs are used as a basis for computing beliefs. Dynamic uncertainty assessment is based on class differentiation. We combine the beliefs of three classifiers: k-Nearest Neighbor (kNN), Naïve Bayesian and Decision Tree. Dempster's rule of combination combines three beliefs to arrive at one final decision. Our experiments with k-fold cross validation show 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. We compare the performance of Dempster's combination (with differentiation-based uncertainty assignment) with those of performance-based linear and majority vote combination models. We study the circumstances under which the evidence combination approach improves classification.

1. Introduction

Medical applications of data mining include prediction of the effectiveness of surgical procedures, medical tests and medications, and discovery of relationships among clinical and pathological data [1]. Clinical databases store large amounts of information about patients and their medical conditions. Data mining techniques applied on these databases discover relationships and patterns which are helpful in studying the progression and the management of diseases [1]. Evaluation may involve prediction or early diagnosis of a disease. In case of diseases like skin cancer, breast cancer and lung cancer early diagnosis is very important as it might help save a patient's life. The aim of this work is to study and apply a formal evidence combination technique for mining medical data for prediction of or screening for a disease. Input data, consisting of feature vectors, is input to three different classifiers. The classifiers we used in this study are kNN (k nearest neighbor) [9], Bayesian [10] and Decision Tree classifiers [10]. Each classifier provides beliefs for each class. These pieces of evidence are then combined to reach a final diagnosis using Dempster's belief combination formula [17]. The experiments are carried on the skin lesion (squamous disease [18]) and breast cancer data [14]. The approach proposed in this paper provides two desirable features: Robustness across multiple data sets with multiple classifiers and management of uncertainty in the presence of unequal error costs. The experiments are carried on dermatology and breast cancer data [14]. Testing is done by k-fold cross validation method with 25% of the data used exclusively as the test set.

In the rest of the paper we first give a brief introduction to the theory of belief functions and evidence. We then describe the computation of belief functions based on classifier outputs in sections 3 and 4. Section 5 discusses how these beliefs are used to compute uncertainty. We then describe our use of the Dempster-Shafer evidence combination approach in the context of the three classifiers. Section 6 describes our experimental evaluation and the results. Section 7 reviews related work and Section 8 concludes the article.

2. Background on the Theory of Evidence

The Mathematical Theory of Evidence is an extension of the probability theory to handle uncertain information [17]. Belief is a measure of a trust or confidence [6, 8, 17]. Let us consider that we have sources of evidence providing various degrees of support for the occurrence of event *A*. Combining all degrees of support for event *A* forms a numerical measure of belief that event *A* occurred. A mathematical function that translates degree of support to belief is known as Belief function. Properties of basic belief m(X) are as follows:

1. $\sum_{X \in \Omega} m(X) = 1$
2. $m(\phi) = 0$ where ϕ is empty. This indicates belief of empty set is always zero.

Belief function for an event *A* can be $Bel(A) = \sum_{X \subseteq A \text{ and } A \in \Omega} m(X)$

Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04)
0-7695-2108-8/04 $ 20.00 © 2004 IEEE

The theory of evidence deals with the evaluation of beliefs from a number of evidences and their combination. For example consider three sources of evidence named E1, E2 and E3. Let the event space be Ω = {A, B, C}. Evidences provide measures for the event space. These measures include belief for each event and uncertainty. Thus measures assigned by evidence E1 are given as $Bel_{E1}(A)$, $Bel_{E1}(B)$, $Bel_{E1}(C)$ and Bel_{E1}(uncertainty). Note that $Bel_{E1}(A) + Bel_{E1}(B) + Bel_{E1}(C) + Bel_{E1}$(uncertainty) = 1. Similar arguments apply to E2 and E3. A decision can be made based on a combination of these beliefs.

In this research we use classifier output to form evidence and a decision such as benign or malignant forms an event. Thus, for instance, Ω = {benign, malignant}. In the following we will illustrate how beliefs for each class and the uncertainty are calculated for each classifier. For the Bayesian classifier, posterior probabilities are used to evaluate basic beliefs. We describe the other two classifiers in more detail.

3. Computing Beliefs with Nearest Neighbors

The k Nearest Neighbor (kNN) classifier considers nearest neighbors as voters. The distance measures evaluated from these neighbors are used to compute beliefs for classes. Distance between test case feature vector and neighbor feature vector is calculated. Let us denote this distance by 'd_s'. This distance is normalized in the range 0 to 1. A fraction of this distance is calculated as d_s/d_{mean} where, d_{mean} is the average distance among the samples belonging to same class and is normalized into the range 0 to 1. Fraction would be greater than 1 if the distance of the test case attribute is more than the average distance for the class and less than 1 otherwise.

To evaluate a distance measure a decreasing function of the distance d_s must be applied. The reason behind this is that as distance between the test case feature and its neighbor feature decreases the possibility that two cases belong to same class increases and the confidence in the event of test sample belonging to same class as the neighbor also increases.

The following distance function is used,

Distance measure = $e^{-\frac{ds}{dmean}}$

where $e^{-\frac{ds}{dmean}} = 1$ when ds = 0

and $\lim_{ds\to\infty} e^{-\frac{ds}{dmean}} = 0$

Thus belief mass of a class is the average of all such distance measures voting for that class. Belief masses for the classes are then normalized so that

$$\sum_{i=1}^{K} m(i) = 1$$

4. Computing Beliefs with a Decision Tree

As explained in the previous section the decision tree classifier builds a decision tree. Association rules can be extracted from this tree. An association rule has support and confidence associated with it.

$$Support = \frac{Number\ of\ records\ with\ A \text{ and } B}{Total\ number\ of\ records}$$

where numerator indicates the number of records with A and B both true.

$$Confidence = \frac{Number\ of\ records\ with\ A \text{ and } B}{Total\ number\ of\ records\ with\ A}$$

Confidence can also be written as $\frac{P(A \cap B)}{P(A)}$

where, $P(A \cap B)$ is the probability of $A \cap B$.

In our context classification process $P(A \cap B)$ forms the probability of the occurrence of feature values with a given class. Here *A* indicates a feature value vector, and *B* indicates a class .

But, $P(A \cap B) = P(A|B) \times P(B)$.

Observing,

$$Confidence = \frac{P(feature\ set \mid class) \times P(class)}{P(feature\ set)}$$

we note that confidence can be used to form basic beliefs of the decision tree classifier.

5. Uncertainty Evaluation

So far we have discussed beliefs are obtained for each class with different classifiers. This section explains how to evaluate uncertainty for each classifier. We use the class differentiation quality as our uncertainty measure [3]. The idea behind this perspective is that the closer the values of beliefs for K classes to each other, the more uncertain the classifier is about its decision. As the beliefs start spreading apart uncertainty starts decreasing. Let uncertainty be denoted as H(U) [3]. Assuming there

are K possible classifications, the distance between the belief values and the value 1/K are evaluated. If all the classes have the same distance then the ambiguity involved in the classification is the highest. If one class shows maximum possible distance then the ambiguity involved is the least. Generalizing from this, a measure of uncertainty can be computed as

$$H(U) = 1 - \frac{K}{K-1}\sum_{i=1}^{K}(m(i) - \frac{1}{K})^2$$

We use this measure to compute uncertainty as Bel(θ)= β H(U) and then normalize the belief values Bel(i) = α m(i) so that

$$\sum_{i=1}^{K} Bel(i) \; + \; Bel(\theta) = \sum_{i=1}^{K} \alpha\, m(i) \; + \; \beta\, H(U)$$

But $$\sum_{i=1}^{K} Bel(i) \; + \; Bel(\theta) = 1$$

And $$\sum_{i=1}^{K} m(i) = 1$$

Thus α = 1- β H(U). In our experiments we have determined a value of 0.3 for β to result in the best classification performance. Dempster's rule of combination deals with these beliefs. Rule assumes that observations are independent and have a non-empty set intersection [6, 17]. Any two beliefs Bel1 and Bel2 with elements Ai and Bi respectively may be combined into a new belief function using Dempster's rule of combination [4]. Let combined belief mass is assigned to Ck, where C is a set of all subsets produced by A∩B. The mathematical representation of the rule is as follows:

$$Bel(Ck) = \frac{\sum_{Ai \cap Bi = Ck;\, Ck \neq \phi} Bel(Ai) \times Bel(Bi)}{1 - \sum_{Aj \cap Bj = \phi} Bel(Aj) \times Bel(Bj)}$$

The combination precedes pair wise. In the first step it combines, for instance, beliefs of k-nearest neighbor classifier (K) and Bayesian classifier (B), and in the second step combines the output of the first step (BK) with the evidence from Decision Tree classifier (D).
Let's assume that the kNN classifier provides beliefs Bel_kNN(B) and Bel_kNN(M), where Bel_kNN indicates belief provided by k-nearest neighbor and B, M are the two classes (benign and malignant) under consideration. Similarly for Bayesian classifier beliefs are given as Bel_Bayes(B) and Bel_Bayes(M). Uncertainties for two classifiers are U_kNN and U_Bayes respectively. Thus matrix under consideration is as follows

Bayes / kNN	Benign Bel_Bayes(B)	Malignant Bel_Bayes(M)	Uncertainty U_Bayes
Benign Bel_kNN(B)	Bel(B)	Bel(Ø)	Bel_kNN(B) x U_Bayes
Malignant Bel_kNN(M)	Bel(Ø)	Bel(M)	Bel_kNN(M) x U_Bayes
Uncertainty U_kNN	Bel_Bayes(B) x U_kNN	Bel_Bayes(M) x U_kNN	Bel(U)

Bel(B) is a belief mass given to class benign. It is evaluated by multiplying benign belief masses of kNN and Bayes, assuming independent evidence sources. Added to this is the product of uncertainty in Bayes and benign belief of kNN, belief for benign of Bayes and uncertainty of kNN. To obtain the combined belief for the hypothesis, all these basic beliefs are summed. Thus,

Bel_comb(B) = Bel_Bayes(B) x Bel_kNN(B)
+ U_Bayes x Bel_kNN(B)
+ Bel_Bayes(B) x U_kNN

This combined belief is then normalized by factor 1 - Σ A ∩ B where A ∩ B= Φ. The underlying assumption for adding the second and the third terms in the numerator is that the uncertainty in an evidence source is a potential support for any hypothesis.

6. Experimental Evaluation

An experimental evaluation was carried out on UCI dermatology and breast cancer datasets [14]. The breast cancer data has a total of 682 instances, each consisting of 10 attributes. All the attributes take values between zero and ten. The classes are benign and malignant and they are denoted as 0 and 1 respectively. 444 records belong to class benign and 238 records belong to class malignant. The dermatology dataset consists of 358 records. The differential diagnosis of erythmato-sqamous is a real problem in dermatology. Types share the clinical features of erythema and scaling, with very little differences [13]. This dataset contains 34 attributes out of which 33 are linear and one is nominal. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. These diseases form the classes for classification. Classification into these classes is denoted as numbers from 0 to 5 respectively. Total records belonging to each class are 112, 61, 72, 49, 52, and 20 respectively. Distribution of datasets among classes is

proportional and thus all the classes are subdivided into 4 different subsets. For testing purposes one subset of each class is kept aside as a test case and the remaining three subsets are used for training purpose.

6.1. Results

The first table below shows the test results of breast cancer dataset in the form of a confusion matrix. Confusion matrices are listed for three individual classifiers as well as the combination. The class denoted by 2 corresponds to indecision. This classification is valuable in contexts where the cost of false negatives is very high. In the case of cancer diagnosis, for instance, misdiagnosing a cancer or its precursor as benign has a very high cost. In such circumstances, it may be preferable to alert an expert with indecision rather than an unconfident and potentially false decision.

kNN				**Bayes**			
	0	1	2		0	1	2
0:	105	2	0	0:	102	5	0
1:	15	48	0	1:	4	59	0
D-Tree				**K+B+D Combination**			
	0	1	2		0	1	2
0:	104	3	0	0:	104	3	0
1:	13	49	0	1:	4	59	0

As shown in this table kNN classifier shows maximum accuracy in classification of records belonging to class 0 (benign). Bayesian classifier shows maximum accuracy in classification of records belonging to class 1 (malignant). Decision tree classifier is less accurate in classifying records of both the classes. As evident from the result set, combination classifier is the most accurate overall. The following tables compare the accuracy of the methods:

Breast Cancer	**kNN (%)**	**Bayesian (%)**	**D-tree (%)**	**K + B + D (%)**
Test 1	90.0	94.7	90.0	95.8
Test 2	90.9	91.5	94.5	93.9
Test 3	90.7	96.5	90.7	95.9
Test 4	96.6	96.6	89.7	97.1
Overall	92	93	91	95.7

In the above table for the breast cancer data set, the overall accuracy of kNN classifier is 92%, Bayesian classifier is 93% and decision tree classifier is 91%. The combination classifier overall accuracy is 95.7%, which is the best overall accuracy.

Skin Lesion	**kNN (%)**	**Bayesian (%)**	**Decision Tree (%)**	**Comb. (%)**
Test 1	43	94.5	96.7	100
Test 2	40.6	91.2	100	100
Test 3	47.2	95.6	94.5	94.5
Test 4	43.5	96.47	98.8	98.8
Test 5	52.8	97.7	89.65	98.8
Test 6	27.7	95.5	90.0	95.5
Overall	42.5	95	94.9	97.9

For the skin lesion data set, the kNN classifier performed surprisingly poorly. Overall accuracy of kNN classifier is 42.5%. Bayesian classifier shows overall accuracy of 95%. Decision tree classifier shows overall classification of 94.9%. Overall combination classifier accuracy is 97.9%. Thus the combination again shows the best overall classification accuracy. The comparison of this table with the previous table highlights a potential benefit of the combination approach: While different classifiers have varying performance on different datasets the combination shows robustness with little sensitivity.

We have compared the results of our Dempster-Shafer combination approach with those of majority vote and linear combination. In both comparisons, the Dempster-Shafer combination achieved the minimum number of misclassifications. The reader is referred to [11] for the details about these comparisons. Comparisons of the Dempster-Shafer combination with other combination techniques such as bagging, boosting and fuzzy logic are future research areas.

7. Related Work

Various classification techniques have been used in the classification of medical data. Examples are wavelets, fractals, artificial neural networks, fuzzy logic, association rules, linear programming and multi-surface separation and nearest neighbors [7, 9,12, 13, 19]. A system proposed in [9] uses Nearest Neighbor, Bayesian classifier and voting feature interval for differential diagnosis of erythmato squamous disease. It provides final diagnosis and explanation from each classifier to the doctors and students. For the detection of breast cancer tumors neural networks and association mining technique was used in [15]. Other techniques for evidence combination include Fuzzy Logic, Neural Networks, General Evidence Processing theory and

Transferable Belief Model [2,4,5,6,16]. The approach introduced in this research work makes use of these individual classifiers. It fuses the results and tries to improve upon the results of the individual approaches. Our experimental evaluation suggested that some classifiers might work better for particular datasets whereas show a poor performance for others. In such cases relying on a single classifier may lead to unacceptable misclassifications. The evidence combination approach facilitates more robust classification over multiple datasets.

8. Conclusion

We have described a method for classifying medical data in the presence of multiple classifiers, uncertainty, and unequal costs of errors. We have demonstrated computation of belief functions and uncertainty values from individual classifiers and combination of evidences through the Dempster-Shafer theory. Class differentiation quality is used for the computation of uncertainties. The combination approach has shown the best classification accuracy across two domains: Breast tumor classification and skin lesion classification. The combination approach remained robust in the presence of fairly different classifier performances. 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. Comparison with other combination methods such as fuzzy logic and neural networks remain as future work. With adaptations, the boosting technique can be applied to classifiers other than the neural network and compared with our approach.

References

[1] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales , M. L. Hage¸ W. Edward Hammond, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse", 1997.

[2] Philippe Smets, Rudolf Kruse: The Transferable Belief Model for Belief Representation. Uncertainty Management in Information Systems 1996: 343-368

[3] Liu Rujie, Yuan Baozong, "A D-S Based Multi-Channel Information Fusion Method Using Classifier's Uncertainty Measurement" Proceedings Of ICSP2000, pp. 1297-1300.

[4] André Ayoun, Philippe Smets: Data association in multi-target detection using the transferable belief model. International Journal of Intelligent Systems 16(10): 1167-1182 (2001).

[5] Lotfi A. Zadeh: Syllogistic Reasoning as a Basis for Combination of Evidence in Expert Systems. IJCAI 1985: 417-419.

[6] Robin R. Murphy, "Dempster-Shafer Theory for Sensor Fusion in Autonomous Mobile Robots" IEEE Transactions on Robotics and Automation, 14:2, 197-206,1998.

[7] M. L. Antonie, O. R. Zaïane, A. Coman, "Application of Data Mining Techniques for Medical Image Classification". MDM/KDD 2001: 94-101

[8] M. Q. Ji, M. M. Marefat and P. J. Lever, "An Evidential Approach for Recognizing Shape Features", Proceedings of IEEE AIA, 1995.

[9] H. A. Güvenir and A. Akkus, Weighted K Nearest Neighbor Classification on Feature Projections in: Proceedings of the Twelfth International Symposium on Computer and Information Sciences (ISCIS XII) S. Kuru, M.U. Caglayan and H.L. Akin (Eds.), Antalya, Turkey, (Oct. 27-29, 1997), 44-51.

[10] C. Borgelt, "A Decision Tree Plug-In for Data Engine", Proc. 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT'98, Aachen, Germany), Vol. 2, pp. 1299-1303, 1998

[11] Mahajani, G. A., Aslandogan Y. A., "Evidence Combination in Medical Data Mining", Technical Report CSE-2003-23, Department of Computer Science and Engineering, University of Texas at Arlington, July 2003.

[12] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[13] W. H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193.

[14] http://www.ics.uci.edu/~mlearn/MLRepository.html

[15] O. R. Zaïane, M. L. Antonie, A. Coman: "Mammography Classification By an Association Rule-based Classifier", Proceedings of MDM/KDD 2002: 62-69.

[16] D. Hall, *Mathematical Techniques in Multisensor Data Fusion*, Artech House, Boston, London 1992.

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

[18] W. Fitzpatrick, *Color Atlas and Synopsis of Clinical Dermatology*, McGraw-Hill 4th Edition 2001.

[19] T. Denoeux. "A k-nearest neighbor classification rule based on Dempster-Shafer theory". IEEE Transactions on Systems, Man and Cybernetics, 25(05):804-813, 1995.