COMBINING CLASSIFIERS FOR BONE FRACTURE DETECTION IN X-RAY IMAGES

Vineta Lai Fun Lum, Wee Kheng Leow, Ying Chen,

Dept. of Computer Science
National University of Singapore
3 Science Drive 2,Singapore 117543
lumlaifu, leowwk, chenying@comp.nus.edu.sg

Tet Sen Howe, Meng Ai Png

Dept. of Orthopaedics
Singapore General Hospital
Outram Road, Singapore 169608
tshowe, mapng@sgh.com.sg

ABSTRACT

In medical applications, sensitivity in detecting medical problems and accuracy of detection are often in conflict. A single classifier usually cannot achieve both high sensitivity and accuracy at the same time. Methods of combining classifiers have been proposed in the literature. This paper presents a study of probabilistic combination methods applied to the detection of bone fractures in x-ray images. Test results show that the effectiveness of a method in improving both accuracy and sensitivity depends on the nature of the method as well as the proportion of positive samples.

1. INTRODUCTION

In medical applications, sensitivity in detecting medical problems and accuracy of the detection (also called specificity) are two important performance measures that are often in conflict. Classifiers that are very sensitive (i.e., have high detection rates) often do so by compromising classification accuracy (i.e., specificity). On the other hand, when the number of negative (i.e., healthy) cases is much large than the number of positive cases (i.e., those with medical problems), a classifier can easily achieve high classification accuracy with very low detection rate.

Such a problem is particularly acute in our application of detecting femur (thigh bone) and radius (arm bone) fractures in x-ray images [1, 2, 3]. Among 432 consecutive cases (i.e., consecutive in date and time at which the x-ray images were taken) of femur images obtained from a local hospital, only about 12% of them contained fractured femurs. For radius images, about 30% of 145 consecutive cases examined contained fractured radius bones. As a result, a single classifier working on a single feature type can often achieve high classification accuracy but very poor fracture detection rate [2, 3]. In [2, 3], a simple voting scheme is used to combine the classifiers to improve both classification accuracy and detection sensitivity.

This research is supported by NMRC/0482/2000.

In this paper, we present a study of the performance of classifier combination in our application context. In particular, the probabilistic combination methods proposed in [4] are tested and compared with the simple voting scheme used in [2, 3]. Test results show that the effectiveness of a method in improving both accuracy and sensitivity depends on both the nature of the method as well as the proportion of negative samples in the test set.

2. RELATED WORK

The first published work on the detection of fractures in x-ray images is that of Tian et al. [1] The method detected femur fractures by computing the angle between the neck axis and shaft axis. Subsequently, Gabor, MRSAR, and gradient intensity were used for fracture detection, and a simple voting scheme was used to combine the individual classifiers that work on single features [2, 3]. Since the individual classifiers tend to complement each other, the combined method improves both the accuracy and sensitivity significantly.

There are three main approaches in combining classifiers. The first approach applies a voting scheme [2, 3, 5, 6]. This is a simple approach that can be used to combine any classifiers. The second approach applies Bayesian theory to derive probabilistic rules such as sum rule and product rule to combine the classifiers [7, 8, 4]. This approach is simple to use but requires that the classifiers output posterior probabilities of classification. Majority voting is a simplification of the probabilistic rule by hardening the posterior probabilities to binary values. The third approach applies boosting techniques such as AdaBoost to weight each classifier according to how well they perform [9, 10]. This approach is more complex. Typically, boosting techniques weight the classifiers according to their classification accuracy only. It is uncertain how to weight the classifiers according to both classification accuracy and detection sensitivity, which are conflicting performance measures. Thus, in this paper, we choose to study probabilistic combination in the context of bone fracture detection.

Fig. 1. Sample x-ray images of (left) healthy and (right) fractured (top) femurs and (bottom) radius.

3. IMAGES AND FEATURES

432 consecutive femur images were obtained from a local public hospital, and were divided randomly into 324 training and 108 testing images. The percentage of fractured cases in the training and testing sets were kept approximately the same (12%). In the training set, 39 femurs were fractured, and in the testing set, 12 were fractured.

145 consecutive wrist images were obtained from the same hospital, and divided randomly into 71 training and 74 testing images. The percentage of fractured cases in the training and testing sets were also kept approximately the same (30%). In the training set, 21 radius bones were fractured whereas 23 were fractured in the testing set. Figure 1 shows sample images containing healthy and fractured bones.

Three types of texture features were extracted from each image, namely, Gabor orientation (GO), Markov Random Field (MRF), and intensity gradient direction (IGD) [2]. Due to differences in gender and age, the same bone of different patients can differ in shape and size. To handle such differences, an adaptive sampling method was employed to sample the features. This method produced feature maps of the same size for the same bone and same feature type [2, 3].

The number of sampling points is inversely proportional to the number of pixels required in a sampling area to accurately extract the features. Gabor features require the most number of pixels and intensity gradient requires the least. Thus, GO maps have the smallest size while IGD maps have the largest size. In addition, the raw femur images are larger than the radius images. So, femur feature maps are considerably larger than radius feature maps. Figure 2 illustrates sample GO and IGD maps of healthy and fractured femurs. MRF maps are not shown because it is difficult to visualize the multi-dimensional vectors of MRF map entries.

The feature maps contain a vector at each map entry.

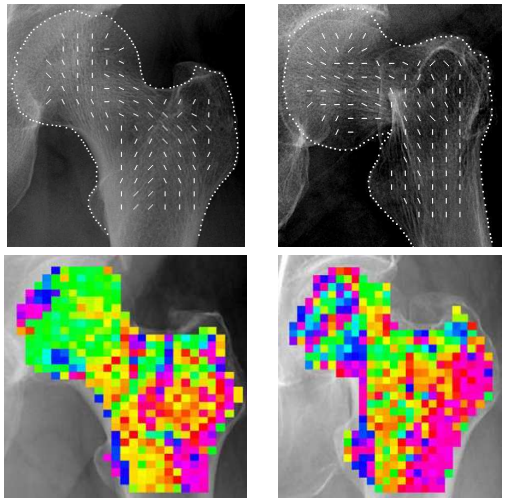

Fig. 2. Feature maps extracted from (left) healthy samples and (right) fractured samples. (top) Gabor orientations are visualized as lines. (bottom) Intensity gradient directions are visualized as colors in the standard color circle.

It is difficult to directly use these vector maps for classification. So, they are converted into scalar maps by computing the scalar differences between the feature maps and the mean feature maps of the healthy training samples [2]. Then, the scalar maps are arranged into feature vectors (containing scalar feature values) for classification.

4. INDIVIDUAL CLASSIFIERS

Gini-SVM [11] was used for classification tests. A systematic method was employed to determine the kernel function and parameter values that produce the best overall performance on the training and testing sets. We found that Gaussian kernel consistently performed better than polynomial kernel. The best parameter values of the Gaussian kernels were found to be 2 for MRF and 0.1 for GO and IGD.

We also tested two other probabilistic classifiers, namely Bayesian and a Matlab implementation of probabilistic SVM [12]. Test results show that Gini-SVM has better overall performance in terms of high accuracy and sensitivity compared to the other two classifiers. So, this paper reports only the test results of Gini-SVM.

Table 1 (top) shows the classification accuracy and sensitivity (i.e., fracture detection rate) of Gini-SVM on femur images. The training performance is nearly perfect for all three types of features. This indicates that the classifiers were well trained. On the testing set, classification with MRF attained the best performance for both accuracy and sensitivity. Classification with IGD had high accuracy but very low sensitivity. This may be because intensity gradient direction does not contain enough information for discriminating between healthy and fractured femurs.

Table 1. Performance of Gini-SVM using single features.

	training set		testing set	
femur	accuracy	sensitivity	accuracy	sensitivity
GO	100.0%	100.0%	89.8%	58.3%
MRF	99.3%	100.0%	98.1%	100.0%
IGD	100.0%	100.0%	90.7%	17.0%
	training set		testing set	
radius	accuracy	sensitivity	accuracy	sensitivity
GO	100.0%	100.0%	90.5%	87.0%
MRF	100.0%	100.0%	86.5%	91.3%
IGD	100.0%	100.0%	96.0%	87.0%

Table 1 (bottom) shows the performance of Gini-SVM on radius images. The classifiers were well trained with perfect performance on the training set. On the testing set, classification with IGD now achieved the highest accuracy but classification with MRF achieved the highest sensitivity.

5. CLASSIFIER COMBINATION

Suppose we have N classifiers based on different feature vectors $\mathbf{x}_i$. Each classifier i measures the posterior probability $P(\omega_j|\mathbf{x}_i)$ of a sample Z belonging to class ω_j, $j \in \{-1, +1\}$ (healthy or fractured) using feature vector $\mathbf{x}_i$. These classifiers can be combined using the rules described in [4]. Five of these combination rules were tested:

Max Rule:

$$\text{assign } Z \rightarrow \omega_k \text{ if } k = \arg\max_j \left[(1-N)P(\omega_j) + N \max_i P(\omega_j|\mathbf{x}_i)\right] \quad (1)$$

Min Rule:

$$\text{assign } Z \rightarrow \omega_k \text{ if } k = \arg\max_j \left[P^{-(R-1)}(\omega_j) \min_i P(\omega_j|\mathbf{x}_i)\right] \quad (2)$$

Product Rule:

$$\text{assign } Z \rightarrow \omega_k \text{ if } k = \arg\max_j \left[P^{-(R-1)}(\omega_j) \prod_i P(\omega_j|\mathbf{x}_i)\right] \quad (3)$$

Sum Rule:

$$\text{assign } Z \rightarrow \omega_k \text{ if } k = \arg\max_j \left[(1-N)P(\omega_j) + \sum_i P(\omega_j|\mathbf{x}_i)\right] \quad (4)$$

Majority Vote Rule:

$$\text{assign } Z \rightarrow \omega_k \text{ if } k = \arg\max_j \sum_i \Delta_{ji} \quad (5)$$

Table 2. Performance of various classifier combinations.

	femur		radius	
	accuracy	sensitivity	accuracy	sensitivity
Max	98.1%	91.7%	95.9%	95.7%
Min	98.1%	91.7%	95.9%	95.7%
Product	20.4%	100.0%	64.9%	100.0%
Sum	22.2%	100.0%	54.1%	95.7%
Majority	91.7%	41.7%	95.9%	91.3%
1-of-3	97.2%	100.0%	85.1%	100.0%
2-of-3	93.5%	41.7%	95.9%	91.3%
3-of-3	89.8%	8.3%	91.9%	73.9%

where

$$\Delta_{ji} = \begin{cases} 1 & \text{if } j = \arg\max_l P(\omega_l|\mathbf{x}_i) \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

In addition, the simple m-of-n rules used in [2, 3] were also tested. That is, a bone was classified as fractured if m of the n classifiers classified it as fractured. So, 1-of-n is the logic OR rule whereas n-of-n is logic AND rule.

Table 2 illustrates the results of classifying the testing samples using various classifier combination rules. For the femur images, max and min rules achieve the highest accuracy of 98.1%. The product, sum, and OR (1-of-3) rules achieve the highest sensitivity of 100.0%. However, the product and sum rules achieve high sensitivity at the expense of very low accuracy. On the other hand, the accuracy of the OR rule (97.2%) is only slightly lower than that of the max and min rules (98.1%). Thus, it can be concluded that for the femur images, the OR rule has the best overall performance of high accuracy and sensitivity.

Interestingly, none of the classifier combination methods significantly outperform the best individual classifier (i.e., using only MRF). The max and min rules achieve the same accuracy as MRF but at the expense of sensitivity. On the other hand, the OR rule achieves the same sensitivity as MRF but at the expense of accuracy.

For the radius images, max, min, majority, and 2-of-3 rules achieve the highest accuracy of 95.9%. Among these four combinations, the max and min rule achieve very high sensitivity of 95.7%, which is only slightly lower than that of the product and OR rule (100.0%). However, the accuracy of the max and min rule is significantly higher than that of the product and OR rule. The majority and 2-of-3 rules achieve the same accuracy as the max and min rule, but lower sensitivity. Compared to individual classifiers, the max and min rules have higher sensitivity than all the individual classifiers, and their accuracy (95.9%) is almost the same as that of IGD (96.0%). So, for the radius images, the max and min rules have the best overall performance of high

accuracy and sensitivity, which is an improvement over the individual classifiers.

For both femur and radius images, the max and min rules consistently achieve the highest accuracy, which is (almost) the same as the accuracy of the best individual classifiers. This shows that the max and min rules consistently optimize the classification accuracy using the posterior probabilities computed by Gini-SVM. They are not as sensitive as the best individual classifier in detecting femur fractures. But, they can detect radius fractures at a higher sensitivity than the best individual classifier. This may be due to the fact that only a small fraction (12%) of the femurs are fractured, whereas a large fraction (30%) of the radius are fractured. Also, the sensitivity of individual classifiers in detecting radius fractures is no where near the perfect sensitivity achieved by MRF in detecting femur fractures. However, it must be commented that this outcome is likely to be an exception rather than the norm because the test results reported here are based on the parameter values that achieve the highest overall performance in both training and testing samples. In real application, MRF alone is not expected to be able to achieve perfect sensitivity. So, classifier combination is still preferred.

The OR rule consistently achieve perfect sensitivity in detecting both femur and radius fractures because it is the least stringent rule in fracture detection. Its accuracy for classifying femur images is marginally lower than that of max and min rules and the best individual classifier. But, its accuracy for classifying radius images is much lower than that of max and min rules. Nevertheless, this result does not diminish the usefulness of OR rule because in real application of screening for possible fractures, one would prefer to have as high a sensitivity as possible while tolerating, say, 10% false alarm rate.

6. CONCLUSION

This paper reported a comprehensive comparison of various classifier combination methods in the context of detecting bone fractures in x-ray images. Test results show that, in general, the max and min rules have the highest accuracy among the various combination methods. The OR rule has higher sensitivity and comparable accuracy compared to max/min rule, especially when the fraction of fractured samples is small.

In real application of screening for possible fractures while tolerating a small amount of false alarm, the OR rule appears to perform better than max/min rule. This is expected to hold when the fraction of fractured cases is small and the various classifiers are complementary in detecting different types of fracture [2, 3]. The OR rule is very useful because it can be applied to standard classifiers that do not output probabilistic measures. For optimizing both sensitivity and accuracy, one may explore a method that combines the strengths of the OR rule and max/min rule.

7. REFERENCES

[1] T. P. Tian, Y. Chen, W. K. Leow, W. Hsu, T. S. Howe, and M. A. Png, "Computing neck-shaft angle of femur for x-ray fracture detection," in *Proc. Int. Conf. on Computer Analysis of Images and Patterns (LNCS 2756)*, 2003, pp. 82–89.

[2] S. E. Lim, Y. Xing, Y. Chen, W. K. Leow, T. S. Howe, and M. A. Png, "Detection of femur and radius fractures in x-ray images," in *Proc. 2nd Int. Conf. on Advances in Medical Signal and Info. Proc.*, 2004.

[3] Y. Chen, W. H. Yap, W. K. Leow, T. S. Howe, and M. A. Png, "Detecting femur fractures by texture analysis of trabeculae," in *Proc. Int. Conf. on Pattern Recognition*, 2004.

[4] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on PAMI*, vol. 20, no. 3, pp. 226–239, 1998.

[5] F. Kimura and M. Shridhar, "Handwritten numerical recognition based on multiple algorithms," *Pattern Recognition*, vol. 24, no. 10, pp. 969–983, 1991.

[6] J. Franke and E. Mandler, "A comparison of two approaches for combining the votes of cooperating classifiers," in *Proc. Int. Conf. on Pattern Recognition*, 1992, vol. 2, pp. 611–614.

[7] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. on PAMI*, vol. 12, no. 10, pp. 993–1001, 1990.

[8] J. Kittler, J. Matas, K. Josson, and M. U. Ramos Sánchez, "Combining evidence in personal identity verification systems," *Pattern Recognition Letters*, pp. 845–852, 1997.

[9] R. Schapire, Y. Freund, P. L. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," in *Proc. Int. Conf. on Machine Learning*, 1997, pp. 322–330.

[10] L. I. Kuncheva, ""Fuzzy" versus "nonfuzzy" in combining classifiers designed by boosting," *IEEE Trans. on Fuzzy Systems*, vol. 11, no. 6, pp. 729–741, 2003.

[11] S. Chakrabartty and G. Cauwenberghs, "Gini-SVM," bach.ece.jhu.edu/svm/ginisvm/.

[12] M. I. Schlesinger and V. Hlavac, "Statistical pattern recognition toolbox," cmp.felk.cvut.cz/~xfrancv/stprtool/.