Hierarchical Classifiers for Detection of Fractures in X-Ray Images

Joshua Congfu He[1], Wee Kheng Leow[1], and Tet Sen Howe[2]

[1] Dept. of Computer Science, National University of Singapore
3 Science Drive 2, Singapore 117543
{hecongfu, leowwk}@comp.nus.edu.sg
[2] Dept. of Orthopaedics, Singapore General Hospital
Outram Road, Singapore 169608
tshowe@sgh.com.sg*

Abstract. Fracture of the bone is a very serious medical condition. In clinical practice, a tired radiologist has been found to miss fracture cases after looking through many images containing healthy bones. Computer detection of fractures can assist the doctors by flagging suspicious cases for closer examinations and thus improve the timeliness and accuracy of their diagnosis. This paper presents a new divide-and-conquer approach for fracture detection by partitioning the problem into smaller sub-problems in SVM's kernel space, and training an SVM to specialize in solving each sub-problem. As the sub-problems are easier to solve than the whole problem, a hierarchy of SVMs performs better than an individual SVM that solves the whole problem. Compared to existing methods, this approach enhances the accuracy and reliability of SVMs.

1 Introduction

Fracture of the bone is a very serious medical condition. According to the International Osteoporosis Foundation [1], 1 in 3 women and 1 in 5 men above age 50 may experience osteoporotic fractures. 30–50% of women and 15–30% of men may suffer osteoporotic fractures in their lifetime. In particular, worldwide incidence of hip fractures can rise from 1.6 million to between 4.5 and 6.3 million by 2050, with more than 50% of all osteoporotic hip fractures occurring in Asia.

In clinical practice, a tired radiologist has been found to miss fracture cases after looking through many images containing healthy bones. Computer detection of fractures can assist the doctors by flagging suspicious cases for closer examinations and directing the doctors attention to suspicious cases. It can thus improve the timeliness and accuracy of their diagnosis.

Computer detection of fractures in x-ray images is a difficult and challenging problem. The femur can fracture in many ways with varying degrees of severity. While severe fractures cause drastic change to the shape of the femur, mild fractures do not change the femur's shape and leave only very subtle signs in the

* This research is supported by NMRC/0482/2000.

W.G. Kropatsch, M. Kampel, and A. Hanbury (Eds.): CAIP 2007, LNCS 4673, pp. 962–969, 2007.

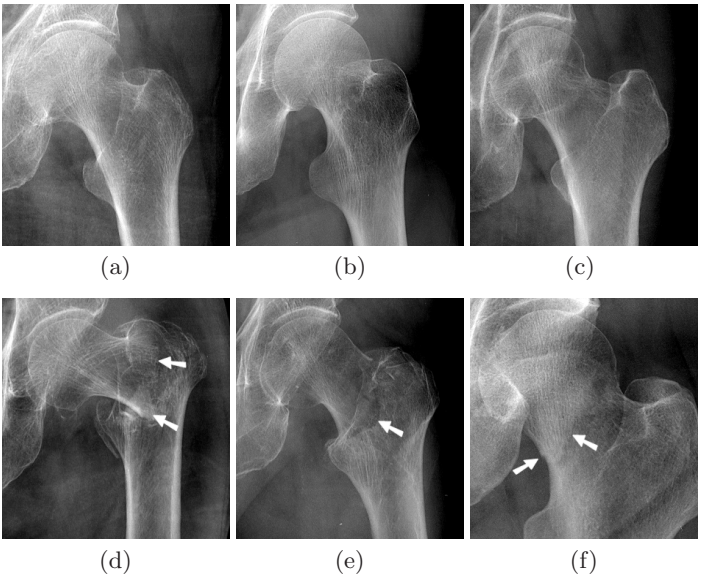

Fig. 1. Healthy and fractured femurs. (a–c) Healthy femurs can have different appearances due to patients' standing postures. (d) Severe fracture changes the femurs' shape. (e, f) Mild fractures leave the femurs' shape unchanged. Arrows indicate fractures.

x-ray images (Fig. 1). There is no single characteristic that can describe all kinds of fractures. X-ray images of healthy femurs also exhibit a significant amount of variation primarily due to the patients' standing postures when the images are taken (Fig. 1).

The unbalanced data problem, i.e., large difference in the proportions of healthy and fractured samples, further compounds to the problem's difficulty. In a consecutive set of x-ray images of femurs (i.e., consecutive in the time that the patients took their x-ray images) that we collected from a local public hospital, only about 12% of them are fractured. When the training set is small, the difficulty becomes even more severe because there may not be enough samples to capture the whole range of possible variations.

This paper presents a new approach for fracture detection by partitioning the classification problem into smaller sub-problems in SVM's kernel space. A hierarchy of SVMs is trained so that each SVM specializes in solving a sub-problem, which is easier to solve than the whole problem. Thus, the hierarchy of SVMs performs better than a single SVM solving the whole problem.

2 Related Work

Tian et al. [2] published the first research work on the detection of fractures in x-ray images by computing the angle between the neck axis and shaft axis. Subsequently, Lim et al. [3] Yap et al. [4] and Lum et al. [5] reported methods

that detect femur fractures based on Gabor, Markov Random Field, and gradient intensity features extracted from the x-ray images. Three SVMs were trained to classify the samples each based on a different feature type. The individual SVM's performance was not very good. By combining the decisions of the three SVMs, the overall accuracy and sensitivity (i.e, fracture detection rate) were improved.

Combination of SVMs is a standard way to obtain a multi-class SVM from binary (two-class) SVMs [6,7,8]. Typically, each constituent SVM is trained to solve a one-vs-one problem, and they are combined using either a tree, a graph, a voting scheme, or other methods. Our method differs from these SVM combination methods in two important ways. Instead of partitioning a k-class problem into many one-vs-one sub-problems, our method partitions a binary (healthy-vs-fractured) problem into several smaller 3-class (healthy, fractured, unknown) problems such that each is handled by a SVM. Moreover, the partitioning is performed based on estimations of the reliability of the SVMs.

3 Hierarchical SVMs

The guiding principle of our approach is *divide and conquer* or *division of labor*. A hierarchy of complementary SVMs are trained to each tackle a different sub-problem of the whole fracture detection problem. A well known divide-and-conquer approach is to first cluster the input samples so that the samples in each cluster are more consistent with each other [9]. Then, a set of classifiers are trained to each classify only the samples in a different cluster. This approach is effective when the training set is large. The more complex the problem, the more clusters are needed to achieve good performance, and the larger the training set needs to be. In our investigation of this approach, we found that a large number of clusters are required to capture the large variations of both healthy and fractured samples. As a result, there are not enough training samples in each cluster to train a classifier.

Instead of partitioning the problem in the feature space, our approach partitions it in SVM's kernel space. This approach has two advantages. First, it is easier to separate the healthy and fractured samples in the SVM's high-dimensional kernel space. Second, the partitioning performed by SVM is optimal.

3.1 Training Algorithm

The training of the hierarchical SVMs is guided by three principles:

1. Samples that can be reliably classified by a higher-level SVM should be handled by it.
2. Samples that cannot be reliably classified by a higher-level SVM should be passed to its child, a lower-level SVM.
3. The performance of a lower-level SVM on the samples passed to it should be better than the performance of its parent on these samples.

The training algorithm begins with the top-level SVM S, which is given two data sets: the training set T and the validation set V. Its main stages are as follows:

1. Train SVM S on training set T.
2. Run S on validation set V to obtain classification error rate $E(S, V)$.
3. Based on $E(S, V)$, select a subset V' of V.
4. Create a new SVM S' at the next level.
5. Find a subset T' of T that can be used to train S' to achieve the performance of $E(S', V') < E(S, V')$ and $E(S', F') \leq E(S, F')$, where F' is the subset of fractured samples in V'. That is, $1 - E(S', F')$ is the sensitivity of S' on V'.
6. If S' cannot achieve the above performance, then stop.
7. Otherwise, set $S \leftarrow S'$, $T \leftarrow T'$, $V \leftarrow V'$, and continue at Step 3.

This algorithm uses probabilistic SVMs, such as Gini-SVMs [10], that produce classification results and probability estimates p. The p value ranges from 0 to 1, with 0.5 corresponding to ambiguous cases located on the decision surface in the kernel space. We shall call the side of the decision surface with $p > 0.5$ the positive side, and the side with $p < 0.5$ the negative side. After running a trained SVM on a sample set, each sample v will be assigned a probability estimate $p(v)$, and the samples can be sorted in increasing order of $p(v)$.

There are two critical stages in the algorithm: Stage 3 and 5, which select appropriate subsets T' and V' to channel to the SVM S' at the next level. In essence, V' defines the sub-problem that S' needs to solve, and T' is the appropriate training set that can train S' to solve the sub-problem. These stages will be discussed in more details in the following sections.

3.2 Selection of Validation Subset

Stage 3 of the training algorithm embodies the first two principles outlined in Section 3.1. It determines a subset V' of V that the SVM S cannot reliably classify. This subset V' is channeled to another SVM in the next level to achieve division of labor. Classification reliability is estimated based on two quantities: (1) the classification error rate $E(S, V)$ and (2) the probability estimates $p(v)$ assigned to samples v in V by S.

Let us compute the cumulative error rate $c^+(p)$ from $p = 1.0$ towards 0.5 for the samples with $p(v) > 0.5$, and $c^-(p)$ from $p = 0$ towards 0.5 for the samples with $p(v) < 0.5$ (Fig. 2(b)). Then, the samples in the range p^+ to $p = 1$ where $c^+(p^+) < E(S, V)$ would have estimated error less than $E(S, V)$; similarly for the samples in the range $p = 0$ to p^- where $c^-(p^-) < E(S, V)$. That is, samples in the tail regions can be reliably classified. Therefore, samples in the middle range from p^- to p^+ should be selected to form the subset V'. In the current implementation, $c^+(p^+) = c^-(p^-) = \varepsilon$ called the *error tolerance*.

3.3 Selection of Training Subset

Stage 5 of the training algorithm embodies the third principle outlined in Section 3.1. Selection of appropriate training subset T' is tricky because SVMs are very strong classifiers. Their accuracy on the training set is often close or equal to 100%. If the method for selecting validation subset is applied directly to the

selection of training subset, very few training samples will be selected and they are not enough for training the lower-level SVM to achieve high performance. On the other hand, if the selection criterion is too loose and almost all the training samples are channeled to the lower-level SVM, then it would be solving the same problem as its parent and there would be no division of labor. So, the goal is to find a subset that is not too large and not too small.

Let $q(u)$ denote the probability estimate of sample u in T. Our method searches for the appropriate T' iteratively:

For q^- from 0 to 0.5 in increments of Δq^-,
  For q^+ from 1 to 0.5 in decrements of Δq^+,
    Set T' to contain all training samples between q^- and q^+.
    Train S' on T' and then test S' on V'.
    If $E(S', V') < E(S, V')$ and $E(S', F') \leq E(S, F')$, then found T' and return with success.
Cannot find desired T'. Return with failure.

The increment Δq^- and decrement Δq^+ are set as fixed proportions of the standard deviations of the distributions of training samples T^+ and T^- in the positive and, respectively, negative side of the decision surface (Fig. 2(a)).

3.4 Testing Algorithm

The same testing algorithm is applied to the validation set during training and testing set at system test. Given a sample v, the following algorithm is applied:

1. For each SVM S from top to bottom of hierarchy,
   (a) Run SVM S on sample v to compute the probability estimate $p(v)$.
   (b) If $p(v) > p^+$, then classify v as healthy and stop.
   (c) If $p(v) < p^-$, then classify v as fractured and stop.
   (d) Otherwise, pass v to the child of S.
2. If with rejection, then classify v as unknown and stop.
3. (Without rejection) If $p(v) > 0.5$, then classify v as healthy and stop.
4. Otherwise, classify v as fractured and stop.

4 Experiments and Discussions

420 consecutive femur images were collected from a local public hospital. They were divided randomly into 200 training, 160 validation, and 60 testing samples. The percentages of fractured samples were kept roughly the same for all three sets at about 12%. Gabor and intensity gradient (IG) features as described in [3,5,4] were extracted as the features for classification.

The following SVM configurations were trained and tested for comparison:

- SVM: Single SVM trained on the training set and tested on the testing set.
- SVM+: Single SVM trained on the combined training and validation set, and tested on the testing set.
- H-SVM: Hierarchical SVM without rejection.
- H-SVM−: Hierarchical SVM with rejection.

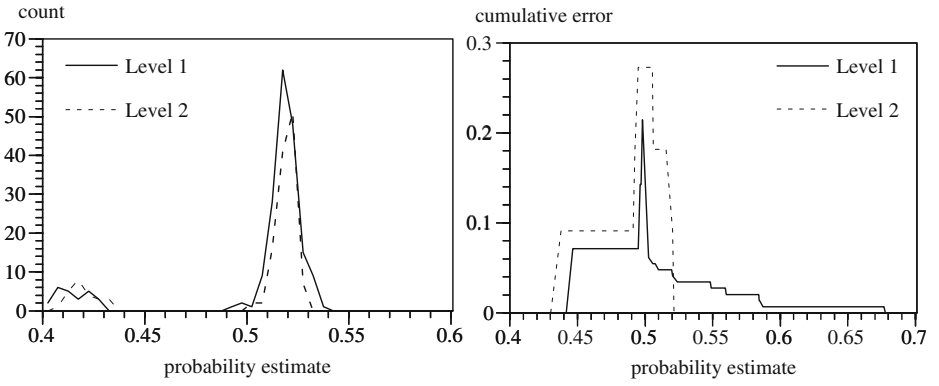

Fig. 2. Performance of SVMs in first two levels of the hierarchy. (a) Distributions of training samples over probability estimates. (b) Cumulative errors on validation sets.

Each of the above configuration was trained to classify Gabor and IG separately.

Figure 2 shows the internal working of H-SVM. The curves on the left and right side of Fig. 2(a) show the distributions of samples on the corresponding sides of the SVMs' decision surfaces. They show that the fracture detection problem is very difficult because all the samples fall within a narrow range of $p = 0.4$ to 0.55, and most of them are on the positive side. Fig. 2(b) shows that the cumulative error at the second level has a narrower spread than that in the first level. That is, errors of the level-2 SVM are focused within a narrower range, indicating that it solves the sub-problem better than its parent.

Figure 3 shows the results of testing H-SVM on the validation and testing sets for both feature types. With very small error tolerance ε, no training and validation samples can be passed down to the lower-level SVMs, reducing H-SVM to a single SVM. With very large ε, most, if not all, the training and validation samples are passed down to the lower-level SVMs defeating the divide-and-conquer strategy. With an appropriate ε, there is a good division of labor. The trends of the error curves for validation set and testing set are similar although their minimum may not coincide at the same ε. In actual application, the error tolerance is selected as the ε that maximizes accuracy (i.e., 1 − error rate) and sensitivity on the validation set.

The trained H-SVMs have a hierarchy of three levels for Gabor and four levels for IG (Table 1). Level-1 SVMs classify more than 70% of the testing samples and pass the remaining samples to the lower-level SVMs. They classify most of the healthy samples and sizable amounts of fractured samples. Therefore, they achieve higher accuracy but lower sensitivity compared to lower-level SVMs, just like single SVMs (Table 2). The samples processed by lower-level SVMs are more balanced, but the discrimination of healthy and fractured cases is still relatively difficult. So, they achieve higher sensitivity at the expense of lower accuracy.

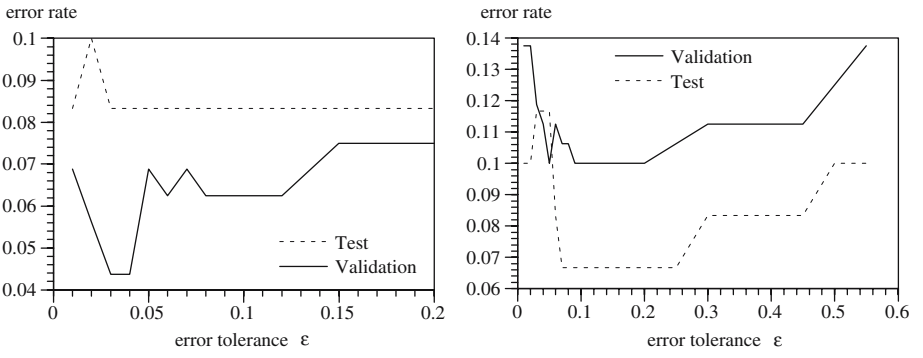

Fig. 3. Test results on different feature types. (a) Gabor. (b) Intensity gradient (IG).

Table 1. Performance of SVMs in the hierarchy. IG: Intensity gradient. accu: accuracy, sens: sensitivity, %class: percentage of testing samples classified by the SVMs in the respective levels, %frac: percentage of fractured testing samples classified.

	Gabor				IG			
level	accu	sens	%class	%frac	accu	sens	%class	%frac
1	95.45%	50.00%	73.33%	28.57%	96.15%	66.67%	86.66%	42.86%
2	90.00%	0.00%	16.67%	14.29%	100%	100%	1.67%	14.29%
3	66.67%	75.00%	10.00%	57.14%	75.00%	100%	6.67%	28.57%
4	—	—	—	—	66.67%	100%	5.00%	14.29%
overall	91.67%	57.14%			93.33%	85.71%		

Table 2 compares the performance of various SVM configurations on the testing set. The performance of SVM+ is better than that of SVM for IG but the converse is true for Gabor. This shows that having more training samples does not always improve the accuracy of single SVM. In comparison, H-SVM can use the training and validation sets optimally to achieve high performance. Its accuracy is as high as the larger accuracy between SVM and SVM+. Its sensitivity is also as high as those of SVM and SVM+ except for IG.

By rejecting uncertain samples, H-SVM− achieves higher accuracy for Gabor and IG at the expense of lower sensitivity compared to H-SVM. The classification results based on Gabor and IG can be combined using a simple OR rule: classify a sample as fractured if it is classified as fractured using either Gabor or IG. For H-SVM−, the combined performance is better than that using single feature. Moreover, its rejection rate drops to 0. With feature combination, H-SVM− achieves a significantly higher accuracy compared to SVM, SVM+, and H-SVM, and the same sensitivity as SVM and H-SVM. In summary, the test results show that the overall performance can be optimized if an SVM can reject samples that it cannot classify reliably, and pass the samples to other SVMs to classify.

Table 2. Test results on different feature types. IG: intensity gradient. Last column is the rejection rate of H-SVM−. OR: Combine Gabor and IG results using OR rule.

		SVM	SVM+	H-SVM	H-SVM−	reject
Gabor	accuracy	91.67%	90.00%	91.67%	94.45%	10.00%
	sensitivity	57.14%	57.14%	57.14%	33.33%	
IG	accuracy	90.00%	93.33%	93.33%	94.83%	3.33%
	sensitivity	85.71%	100%	85.71%	83.33%	
OR	accuracy	86.67%	88.33%	90.00%	93.33%	0.00%
	sensitivity	85.71%	100%	85.71%	85.71%	

5 Conclusion

This paper presents a new divide-and-conquer approach for fracture detection by partitioning the problem in the kernel space of the SVM into smaller sub-problems, and training an SVM to specialize in solving a sub-problem. Each sub-problem is easier to solve than the whole problem. The training scheme ensures that lower-level SVMs always complement the performance of higher-level SVMs. As a result, the hierarchy of SVMs performs better than an individual SVM solving the whole problem. Compared to existing methods, this approach can enhance the accuracy and reliability of the SVMs.

References

1. IOF: Facts and statistics about osteoporosis and its impact. International Osteoporosis Foundation (2007), `www.iofbonehealth.org/facts-and-statistics.html`
2. Tian, T.P., Chen, Y., Leow, W.K., Hsu, W., Howe, T.S., Png, M.A.: Computing neck-shaft angle of femur for x-ray fracture detection. In: Petkov, N., Westenberg, M.A. (eds.) CAIP 2003. LNCS, vol. 2756, pp. 82–89. Springer, Heidelberg (2003)
3. Lim, S.E., Xing, Y., Chen, Y., Leow, W.K., Howe, T.S., Png, M.A.: Detection of femur and radius fractures in x-ray images. In: Proc. 2nd Int. Conf. on Advances in Medical Signal and Info. Proc. (2004)
4. Yap, W.H., Chen, Y., Leow, W.K., Howe, T.S., Png, M.A.: Detecting femur fractures by texture analysis of trabeculae. In: Proc. ICPR (2004)
5. Lum, V.L.F., Leow, W.K., Chen, Y., Howe, T.S., Png, M.A.: Combining classifiers for bone fracture detection in x-ray images. In: Proc. ICIP (2005)
6. Kreßel, U.: Pairwise classification and support vector machines. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods Support Vector Learning, pp. 255–268. MIT Press, Cambridge (1999)
7. Platt, J., Cristianini, N., Shawe-Taylor, J.: Large margin DAGs for multiclass classification. In: Solla, S.A., Leen, T.K., Muller, K.R. (eds.) Advances in Neural Information Processing Systems, MIT Press, Cambridge (2000)
8. Hsu, C.W., Lin, C.J.: A comparison of methods for multi-class support vector machines. IEEE Trans. on Neural Networks 13, 415–425 (2002)
9. Li, R., Leow, W.K.: From region features to semantic labels: A probabilistic approach. In: Proc Int. Conf. on Multimedia Modeling, pp. 402–420 (2003)
10. Chakrabartty, S., Cauwenberghs, G.: Gini-SVM. (bach.ece.jhu.edu/svm/ginisvm/)