

Course on Introduction to Medical Imaging and Analysis Softwares
Professor Debdoot Sheet
Department of Electrical Engineering
Indian Institute of Technology Kharagpur
Module No 2
Lecture 10-Systematic Evaluation and Validation

So welcome to this very important lecture in perspective of trying to prove your methods which you have developed till now and this is called as systematic evaluation and validation. So this is a place where we are going to learn about how you are going to benchmark your algorithms which means to say as to how good is your method performing as compared to other methods so if you are looking in perspective of developing something new.

If you are looking in perspective of actually trying out what somebody else has worked out already on then this is a particular chapter you will need to understand in order to evaluate if there are 7 different methods available to solve one single problem and which of these methods is most suited to solve your problem and what are the tradeoffs you can always achieve. So maybe you can make it faster compromising certain part on performance or maybe the other way round that you do not want to be faster, you have enough of time to spare on it but you actually wanted to be the most accurate as much as possible. So we will be discussing about them one by one.

(Refer Slide Time: 1:21)

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Contents

- Datasets and baselines
- Prospective vs. retrospective experiments
- Bias and variance
- Sample sufficiency
- Evaluating segmentation
- Evaluating classification
- Receiver operating characteristics
- Folded Cross-validation

Systematic Evaluation and Validation (Debdoot Sheet) 2

the way this lecture is actually organized is I would be initially speaking about what is as per definition datasets and what are baselines then I would be telling you about retrospective versus Prospective experiments and what they mean. We will also be discussing about the Bias variance problem and the Bias variance tradeoff which we often speak about in some kind of a pattern recognition machine learning or computer vision task and what the perspective means for medical image analysis as well.

Then we will be speaking about the sample sufficiency issues and what when do you call something as a sufficient sample. Following that we would be entering into certain very standard evaluation methodologies, the first one will be for evaluating segmentation algorithms and segmentation performances, following that we would be entering into classification evaluations. So classification evaluation can also be used for segmentation problems if you are looking at segmentation as a pixel wise classification problem.

Otherwise you can also be taking them as image level classification problems or volume level classification problems or even case level classification problems. From there we enter into a very standard way of evaluating your performances of algorithms in any kind of a learning based task and that is called as a receiver operating characteristics. So once we have all of this ready that you have you know about what datasets are what your baselines are where your methods are standing with respect to others in terms of a very standardized experiment the major thing which comes down now is you need to cross validate.

And we will be learning about how to do a folded cross validation which basically means that can you keep a part of your data as blinded and not use it during training and see how the performance comes. And now can you actually do it repetitively or iteratively over a cycle of procedures so that you know what is the best case performance, worst case performance and average performance of your algorithm and it is actually put into a field level deployment.

(Refer Slide Time: 3:26)

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Datasets and Baselines

Playground	Prior Record Holders
<ul style="list-style-type: none">• Standard evaluation<ul style="list-style-type: none">– Anonymized– Same modality or set of modalities– Cohort collected<ul style="list-style-type: none">• Similar disease• Similar genetic makeup• Similar age-group and demography– Equal number of subjects per pathological group<ul style="list-style-type: none">• Control / normal may be absent• Follow-up / prognosis data<ul style="list-style-type: none">– Similar number for all cases used	<ul style="list-style-type: none">• Literature reviews<ul style="list-style-type: none">– Performance of methods employing same dataset– Re-implementation of other methods on standard dataset– Same set of metrics used for evaluation<ul style="list-style-type: none">• Comparing computation times - similar hardware platform (acceleration if any) to be used

Systematic Evaluation and Validation [Deebot Sheet] 3

So let us get started, the first point is about Datasets and Baselines, now as I put this one is it can very categorically be said as a playground and all the prior record holders. Now imagine any of your algorithm development task is actually a one single player game, it is no more a team player game even if you were say team player game them we would rather call this as a world cup or some sort of a team playing games over there, otherwise till consider this to be an Olympics of single events.

Now over there your playground is basically where you are going to play the game you are going to do. So obviously you are not going to do some sort of a pole vault in a swimming pool that is quite out of way so you need to understand as to where what is your playground and that is called the dataset where we are going to work on. On the similar way, if you are actually somebody who is competing to be the best swimmer you are not going to look into timing performances across pole vaulters or across javelin throwers.

So that is the point where you need to understand as to what baselines are which we are going to use. Let us get into the first part of it which is your playground or the dataset. Now if you look into the dataset part over there then we have very standard paradigms of evaluation. So if you look into the dataset part over here we have very standard paradigms of evaluation, the first and foremost thing to be guaranteed for medical experiments is that the data needs to be anonymous, you cannot compromise on leaking out the patient's information at any point of time.

Now information is obviously the name or some entity by which the particular subject whose data you are using can be identified. For all practical purposes of research and for all ethical concerns we cannot divulge out the patient information in anyway and for that particular reason you would see that most of the data sets which are publicly available they will never divulge out this information, the patient field is always kept as blank.

The only thing you might ever get is basically the age of the patient which is actually a clinical identifier, the gender of the patient which is again a clinical identifier and sometimes you would be getting down extra information like the race of the patient, the kind of economic background they come from the education background or the patient kind of jobs they are doing. So these are extra metadata which play a lot of significant part in terms of developing some sort of a medical diagnostics system.

The next one is that you need to choose the same set of modalities or the same modality. So if I am trying to do an experiment in which I want to segment out say the left ventricle of the heart, now I cannot say that I will be segmenting left ventricle of the heart and trying to prove its performance using 5 images of ultrasounds, 7 images from MR in T1, 3 images from MR in T2 and say 8 images from a CT angiography system.

Now this is way wrong, you need to be either focusing on one of the modalities or say if you are doing a multi modal then for all of your patients if we say you have about 35 patients over there on which so for our purposes w will call them as subjects because we do not know whether they are actually patients or not they might be healthy individuals as well.

So if you have 35 subjects as well, you should be having all of the modalities for all of them, the same set of modalities or maybe the only modality on all of them, this is how you need to prove and you need to very conscious about that. You cannot compare apples and oranges so you cannot compare performances of segmentation on ultrasound with performance of segmentation on CT images. The next one is that you need to have a Cohort from where you are going to collect all the data and this cohort means that you need to have similar set of diseases.

So if you are trying to quantify certain cardiovascular diseases then you will need to necessarily know that all of these subjects they come with a prior history of cardiovascular problems or they

are perfectly normal as far as cardiovascular problems are concerned, but then if you want to look at segmentation for certain disease indicators or you are segmenting a particular lesion of the brain and all the subjects which you are getting down over here on your MR scan of the brain on which you want to segment a lesion never had any sort of a brain problem or they never had a problem in terms where a lesion would form in the brain, then your cohort is actually flawed. You cannot use that data for doing over there.

The next one is that they need to come down from a similar genetic makeup. Now all of us by now in this global age we do understand that genetic makeup plays a significant role in how our bodies are formed, how tall we are, how wide we are, how what is our weights and everything in fact the size of the skull, the shape of the skull in which the whole brain is also encapsulated varies from different genetic groups to different genetic groups.

So you need to be very cautious about using a cohort from a single genetic makeup or even if you are doing it from say three different genetic makeups, so say let us consider that you are taking the Indian, you are taking an African genetic southern African genetic makeup and you are again taking a European genetic makeup then be very sure that you are actually going to take equal proportion of people from all of them and they should also be balanced in terms of the number of the proportion of genders, so you cannot take all females from the Indian population and all males from the African population and then try to come up with a disease cohort summary.

So you need to have both males and females. If it is spread over a span of age group which means that this disease can exist from 10 year old to 100 year old then you should be having representative population on each of them, you cannot say that my Indian population just has people from 10 year old to 40 years old and my African population has all people from 50 years old to 100 years old and now I want to compare whether the effect and you cannot generalize and say that with whatever findings you are getting down on the 50 years old to 100 years old on African population you cannot bring the same thing and say for the Indian population of 50 years old to 100 years old because for the Indian population you just had for the group of 10 years old to 40 years old, now this is what you need to keep in mind.

The next part is that the demography also has to be kept in concern, so these are major things you need to know when you are planning a cohort. For most of the publicly available datasets these are actually sorted out because when data collection goes through an institutional regulatory approval, this complete experiment design of how the data will be collected and how it will be annotated and stored is actually reviewed much before the data collection starts and even after the data collection before the data is actually released for public use and said to satisfy a certain quality based on which algorithms can be evaluated.

It is again internally audited in order to check and verify whether everything which was promised and was expected by the committee of or regulatory approvals has been satisfied before the data is released down. The next part is that you need to have equal number of subjects for pathological group which means that basically you if you are looking for certain sort of a cancerous lesion, now cancers can be of different categories there can be type 0 which is basically when no cancer has been found to a type 1 where it is not yet a cancer it is some sort of a tumor but in a benign form and most and least likely to ever become a cancer then there can be type 2 which is it is exhibiting some sort of growth but that is not exactly a cancer.

Then there is type 3 which is a border line case which means that it was sort of type 2 maybe a few instances of time ago which is a few days ago or few weeks ago, but now it is sort of showing a progressing behavior towards type 4 which is a border line case for start of a cancer and maybe till type 5 which is case where you definitely have a malignancy proven down and type 6 where it is actually infiltrating out and going to infect other organs as well and is the most critical case.

Now if you are looking at each of these classes it is expected that you have data for each of these class. Now you cannot say you do not have data for type 0 it is okay because normal people who do not have cancers might not turn up for it, okay. But if you do not have data for type 1 you do not have data for type 2 you cannot generalize your algorithm performance for type 1 and type 2. You will always be reporting that you are able to only validate performance for type 3 and type 4 type 5 lesions for which you have your images available to you.

So this is one thing you have to keep in mind you cannot say that I have a universal lesion segmentation algorithm of some sort when you do not even have data sets from type 1 and type 2

available to you. So this is another major concern to be kept in mind although in medical images we do understand that getting normal is a problem and that is always kept inside the reservation saying that without normals just for the disease categories we are definitely able to prove.

Now, the other part is that, if you have access to a follow up data or prognosis data or your complete cohort, use it only in that case. Suppose you do not have follow up data for all of your subjects enrolled in the study or all the subjects from who's data you are using that case do not use it. So follow up basically means that the subject came in for the first time, you got a certain sort of a diagnosis you have the imaging data and everything available to you, you are doing your segmentations and everything now the subject comes down for the second time you cannot consider this to be a second case anymore this is actually a follow up of the first case itself.

So your sample is not increasing over here, basically your temporal sampling of the observations are increasing. So your number of observations is more but number of samples is not more, this is what you need to keep in mind and use a follow up only if you have follow up available for all the or significant sizeable portion of your population. Say you have some 80 subjects and you have 2 different classes, so 40 subjects per class, now for the first class you never have a follow up and for the second class you have just have keep on having follow up.

Now if you want to report follow up you can you need to very categorically say that I am reporting follow up only for this category to of subjects for whom follow up was recommended and out of these 40 subjects over there be I am reporting only for 30 who actually turned up for follow up and this is the observations on the differential side of it. So this is how reporting has to be very crisp and clear when you are mentioning at some point.

Now from there we move onto as to what is about the prior record holders and what we would mean to say in terms of what records we need to break. So over there you would clearly look that you will be getting about the prior record holders from your literature reviews. So in this literature reviews basically all the previous methods all the previous papers which were reporting something which is related to the work over there.

Now in case, you see that those methods where reporting on some other dataset or their own dataset which is not publicly available. And it is generally suggested that you re-implement that

method on your standard dataset and report it because two different datasets and reporting performance on them and trying to compare is basically farce over here, whereas what you are trying to in effect do is that say I am in terms of cricket let us do it something that I am trying to prove team A's capability of playing a 20-20 match to team B's capability of playing a test series.

Now that is way wrong because everything has a different game plan everybody has a different games span over which they are doing somebody might be a good sprinter performer on a shorter period of time, somebody might be somebody who can play on for a longer period of time. Now this is wrong way of proving, instead of that if you really want to compare team A and team B in playing 20-20 you need to make both of them play 20-20 or if you want to make each of them find out the performance on a test then play one against the other on a test match itself, you are not going to compare one team for playing 20-20 to the other team for playing a test match.

Now this what you need to keep in mind over here that you need to re-implement the method and use it over here. Most likely you would be getting an open source implementation for the algorithms which makes it much more easier. Now from there you would also need to specify if you are using the same set of matrix for evaluation, so if you are using accuracy for benchmarking everything use accuracy, do not just branch of at some point of time say they at algorithm b and c and d have reported sensitivity so we are just going to report sensitivity and compare them.

So that is also a wrong way of doing it. We come down to what they mean and how they will be impacting your total performances over there. Including that always mention about what is the performance and what was the time taken down because in this particular field we are very cautious about how much of time you are consuming in order to do a certain experiment and how you would be validating that. So from there we come down to where to get datasets.

(Refer Slide Time: 17:05)

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Datasets

Grand Challenges in Biomedical Image Analysis

All Challenges

Here is an overview of all challenges that have been organized within the area of medical image analysis that we are aware of. If you know any other challenge, please leave a message in the [forum](#).

Showing 130 projects of 130

Filter by: Open for submissions (94) Data download (94) Archived (130)

2017

Workshop: Apr 18, 2017
Associated with: [ICB 2017](#)
Hosted on: [grand-challenges.org](#) Associated with: [ICM 2017](#)

CAMELYON17
Automated detection of lymph node metastases in histological images of prostate sections. The task has high relevance and would normally require extensive microscopic assessment by pathologists.

FROSTATEx
Diagnostic classification of clinically significant prostate lesions using quantitative image analysis methods.

Systematic Evaluation and Validation [Deebot Sheet] 4

Now as I had mentioned on the first lecture itself is a very good place is actually going down to grand-challenges dot org and trying to find out from there what are the existing challenges and if there is something new which has come up over there or not. Beyond that you also have a lot of public datasets which we would be listing down on the (())(17:24) as well so that you can get access, you can have a readymade lookup access to them as well.

(Refer Slide Time: 17:31)

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Benchmarking on a Dataset

DRIVE: Results Browser

Next Prev Go to 1 Magnification factor: 0.2 display soft classification when available

Display the following: input gold standard human observer Chaudhuri Jiang Niemeijer Perez Staal Zana

Results for case 1.

Displayed	Sensitivity	Specificity	Accuracy	Az
1. Input				
2. Gold standard				
3. Chaudhuri	0.276	0.997	0.903	0.950
4. Jiang	0.714	0.969	0.918	
5. Niemeijer	0.719	0.972	0.939	0.944
6. Perez	0.796	0.961	0.939	
7. Staal	0.778	0.971	0.940	0.967
8. Zana	0.773	0.975	0.949	0.942

<http://www.isi.uu.nl/Research/Databases/DRIVE/browser.php>

Systematic Evaluation and Validation [Deebot Sheet] 5

Now benchmarking typically what it would be looking like is something like this. So I have taken down these results from the drive dataset or retinal vessel attraction. And they have a very

intuitive method in which what they do is you have the original image, you have a ground truth and then performances of different methods, so you have small thumbnails over here and then the accuracy sensitivity, specificity, accuracy and area under the roc curve which is reported for all of them.

And this is a typical way in which benchmarking is done, so this is an image to image base now for all the images if you take it together then you can consolidate it and always give down an average value coming down over the whole dataset for this particular one and all of these are different methods and by the authors who are present over here. So this is a way in which benchmarking is generally expected when you are doing it over a dataset and it is expected that if you are writing on a paper or report some place or consolidating your own performances then you report in similar way over there, you put down what where are is because do not expect that the person who is reading it or evaluating it is an expert or can readily recall what performances where.

So you are expected that you put down everything on a single piece of paper so that it becomes very much evident for somebody to track as to whether you are good or how even if you are bad then how bad and is there some sort a trade off which you are able to achieve.

(Refer Slide Time: 18:54)

The slide is titled "Experiments" and is presented by the Indian Institute of Technology Kharagpur, Department of Electrical Engineering. It compares two experimental approaches: Prospective and Retrospective. The Prospective approach involves planning data collection before an event occurs, often used for testing hypotheses, and has pros like class balance and cons like regulatory approvals and cost. The Retrospective approach involves collecting data as an event occurs, often used to form hypotheses from observations, and has pros like lower cost and regulatory ease, but a con of class imbalance.

Prospective	Retrospective
<ul style="list-style-type: none">• Plan data collection before event occurs<ul style="list-style-type: none">– Pharmaceutical / drug trials– Controlled animal model trials• Used to test a certain hypothesis• Pros<ul style="list-style-type: none">– Class balance• Cons<ul style="list-style-type: none">– Regulatory approvals– Costly to perform	<ul style="list-style-type: none">• Data collected as event occurs<ul style="list-style-type: none">– Epidemic data– Imaging modality efficacy studies• Used to form a hypothesis from observations• Pros<ul style="list-style-type: none">– Less expensive and generally free of cost– Regulatory approvals hassle free• Cons<ul style="list-style-type: none">– Class imbalance

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Systematic Evaluation and Validation [Debdoot Sheet] 6

Now from there I would be speaking about what experiments are. So typically we have two types of data collections within experiments, one of them is called as a prospective, the other one is called as retrospective. Now in prospective what you do is as the name suggest you basically plan for the data collection before the event occurs. So this is what is done by pharmaceutical companies for drug trials and you typically use a controlled animal model for this sort of trials. There are humans who also enroll for similar kind of trials over there but they are much more serious affairs and come much later after animal trials and all sort of regulatory approvals and ethical clearances.

Now here the idea is basically to test a certain sort of hypothesis. So say you are coming up with a hypothesis that given I am injecting certain sort of a drug to a person with this kind of a grade of tumor then the tumor is going to go down. So there will be animals on whom this same sort of tumors will be created. So say 100 animals on which the tumors is created. And another 100 animals on whom the tumor is not there. And now what we what they do is they keep on injecting the drug on each of them and they do a follow up study.

Now on the group of animals on whom the drug has been injected if there is a hypothesis that the lesion is going to shrink, then you will be observing for the shrinkage and lesion. Now for image analysis where the job comes in that the lesion would obviously be imaged by somehow, so maybe a CT or NMR and your job is basically segment it out and do volume estimations over time and find out. So this is definitely a follow up experiment which you are going to get down over there. The number of samples cases you have for healthy and the number for the tumor model is always same, so you have a data balance over here completely.

And on the other side of it what you will be looking is that whether the drug is inducing some sort of a side effect or secondary lesions in healthy subjects as well. So this is what you need to keep in mind during evaluations. Now the good thing about this experiment is that you have a class balance because now you can ensure how many I need for the normal one, how many I need for the abnormal one, the bad thing is that this definitely requires a lot of regulatory approvals because you are subjecting some living beings onto your experiments and making them skip boards. So you need to be very cautious about that.

And next is obviously they are very costly to perform because your keeping everybody informed and the whole data collection and everything is from your own pockets which goes into it. But obviously this is the basis for doing pharmaceutical trials and how new drugs come into testing and then eventually into market. The next one is a retrospective study and this is when the data is collected as the event occurs, so one common example is Epidemic data. So there is epidemic breakout or something and the health offices are constantly recording that data.

So they are not planning as to when the epidemic is going to occur, but as it occurred as there was a (())(21:57) the disease the data is just kept on being collected over there. So this is where you have a you will not always have imaging modalities but in a lot of cases you will be having imaging modality efficacy studies which go over here. So which means that over the long run so say there is a center or there is a particular machine whose quality is to be evaluated so whatever scans are being done on the machine everything is recorded over there and then people are going to use this at a later point of time in order to evaluate whether the whole study was good or not or the machine was performing good or not.

Now, over here what we do is basically we need to form a hypothesis from the observations itself. So here the hypothesis previously not note we just observed certain things and then sort of create a model in order to form a hypothesis. Now the good thing about this kind of an experiment is that it is obviously less expensive and generally free of cost because it is just in a treatment protocol when a or a diagnosis protocol when the images or data are getting acquired you are just going to make use of this information for your research.

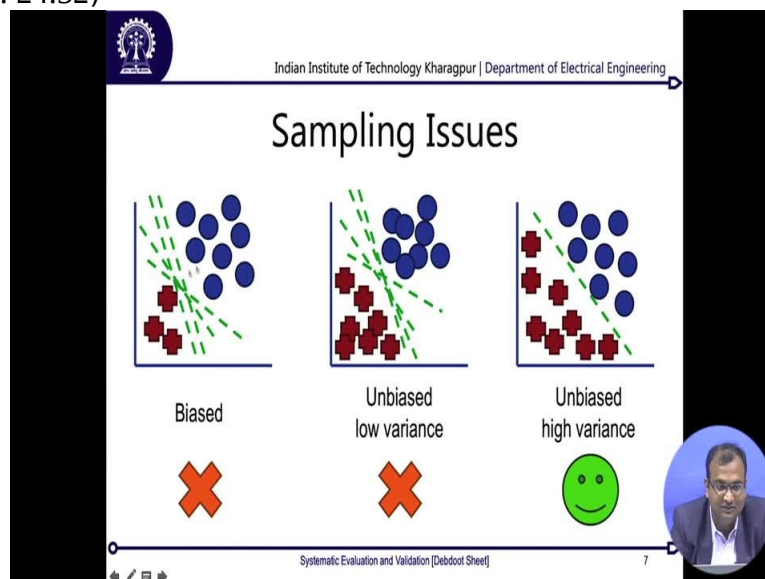
So regulatory approvals are also generally hassle free because the event has happened the data has been recorded, you are just going to get a regulatory approval for ironimizing the data and using it for proving down a particular hypothesis or for basically designing a particular hypothesis by looking at the data. Now the cons is this will have severe class imbalance and the reason being that say there is a particular say we again get back to the cancer example from class 0 or type 0 of cancer till type 6 of cancer.

Now type 0 most subjects will not be coming down for imaging study because they do not have a lesion, type 1, type 2 is what you will be having a lot of patients and in in fact like about 80 to 85 percent of subjects will be just type 1 and type 2. And on the other side of it type 6 which is the

most crucial one which you would like to basically be very sensitive to will have maybe 2 or 3 patients just coming down for those.

Now if you look at the relative distribution of each of these classes it is never balanced, so some classes would be very heavy, some classes would be very low and there are chances that whatever classifiers so learning in just you are going to develop, you are going to get very much biased towards this particular one which is very high. Now we come down to this issue on the sampling in the next slide where I would be explaining them in much more details as to what happens.

(Refer Slide Time: 24:32)



Now about the sampling issue which I was telling you, now one particular case which we call as a bias sampling. So let us take just two categories of problems so I have my red squares and red plus symbols and I have my blue circles over here. Now these and then these green lines are basically my separation margins in a classifier. Now what I wanted to say was you will be seeing that there are multiple of these separation margins over here and the reason is that based on how your learning system is initialising itself that you can get any one of these.

So there can be multiple classification margins and separation margins as we had learnt in the previous lectures on simple classifications. Now what happens over there is that as you would be having an biased class which means that there are more number of samples from blue and say

this is what I said was benign cases or type 1 and type 2 okay and you have less number of samples from malignant classes now the problem which comes over here is that you will be heavily biased towards making a decision in favor of the benign and you will be mistaken for malignant ones and that is when this margin shifts a lot far away from here so anything which comes in this border zones they are most likely to be may called as benign and less likely to be called as a malignant, okay.

Now the other problem is that you also have sort of a variance problem in the second case. In this second case what happens is that you will see that the number of samples you have for benign and malignant is sort of same almost same number of samples. But the problem is that they are not quite spread off in this part which means that there is lot of grey zone over here. Now if you look at this total feature space lot of this space is basically where you never had any samples come on.

Now if there are samples coming on from here, you are just basically left at like the mercy of whatever classifier line came down over there and there is no assurity that you are doing it with a very confident decision. Now this is a problem which happens with a unbiased but a dataset which has a low variance, whereas the best one is actually an unbiased and high variance dataset which means that your dataset is quite spread across the whole of feature space such that you are getting a classification margin which is a unique classification margin and the chances that it makes error is actually much lower.

Now how to do you ensure this is a major question. I mean until possible you have the classification margin you can never ensure it. But at the point of data collection if you can ensure that you have samples from at least one sample from every single manifestation coming down and you have some sort of a balancing between the number of positive samples and negative samples, you should be able to get down a way in which you can always have a unbiased estimator with a very high variance on your samples sets being created.

So this is an issue which you will have to really really take care of by yourself, but the problem is that it is not so easy I mean it is much easier said than to be done because samples sufficiency for medical images is actually a myth it is a serious myth.

(Refer Slide Time: 27:46)

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Ensuring Sample Sufficiency

- Sample sufficiency?
 - Myth for medical image datasets
 - Normal or healthy
 - More cases and samples
 - Abnormal or diseases
 - Rarer a diseases – lesser the samples
 - Require high performance for rare diseases
- Solution
 - Data augmentation during training
 - How?
 - Replicate samples for the weaker class
 - Use rotations, affine transformation, etc.
 - Restrict use of warping on images or applying intensity transformations.

Systematic Evaluation and Validation (Deebot Sheet) 8

Now you will have more cases and samples for normal or healthy or say all the benign cases. And for abnormal diseases the rarer a diseases the lesser is the number of samples you will get and the problem is that you will always need this performance margin to be highest for this rarer disease.

Now the solution is actually in order to do something called as data augmentation. Now this is a standard technique used for a long period of time in computer vision practices and the idea was much more simpler in computer vision because you have a lot of natural languages natural scenes which you are using over there and from these natural scenes you can actually augment it out synthetically augment out something but then for medical imaging you need to be very cautious as to how you are augmenting.

Now, first and foremost how you need to do is that you need to replicate samples for the weaker class, okay. So replicating samples are not going to just solve your problem because what you would be doing is on the bias variance space if you look at it you will still be having a much lower variance you are not going to increase the variance you are just compensating for the bias by replicating samples. Now in order to compensate for the variance as well what you can do is you can introduce some sort of rotation and affine transformation which is you can do a bit of a skewing, scaling or you can rotate the lesions or image appearances a bit within certain sort of a tolerance limit.

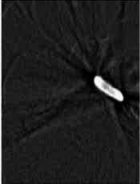
So obviously a CT scan of an upright person and a top down person is very different so you cannot do that much of a 180 degree rotation but maybe within a 5 degrees or 10 degrees of tolerable limit you can do some amount of rotations over there. But definitely restrict the use of warping or applying intensity transformation because these are very disease specific and lesion specific. So by applying a warping you are basically a lesion which is oval and you suddenly warp in a weird way and it becomes jiggled in shape so that is is no more a benign lesion in any way. So that becomes very close to a malignant lesion.

So this is where your augmentation is going to introduce errors into your learning system. The other one is that you if you are applying an intensity transformations these are factors of tissue energy interactions that we had studied in the previous lectures. So you are going to again play around and some sort of introduce errors into this tissue energy interaction which is very specific for diseases and modalities. So you cannot use these ones. Now this these are just certain suggestions, there is definitely no defined way of doing it and this are how we have been able to solve a lot of problems till now.


(Refer Slide Time: 30:22)

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Segmentation




ground truth



$g(\mathbf{x})=\{A\}$

delineated object



$f(\mathbf{x})=\{B\}$

$$O = \frac{|B \cap A^c|}{|A|} \quad U = \frac{|A \cap B^c|}{|A|} \quad D = \frac{2|A \cap B|}{|A| + |B|} \quad J = \frac{|A \cap B|}{|A \cup B|}$$

Oversegmentation Undersegmentation Dice coefficient Jaccard coefficient

Systematic Evaluation and Validation [Debdoot Sheet] 9

Now from there I come down to very simple matrix so say that you have a segmentation problem so this was an object which you had to segment and there is an some sort of lesion in an MR which you were wanted to segment. Now there is obviously a ground truth which is available we

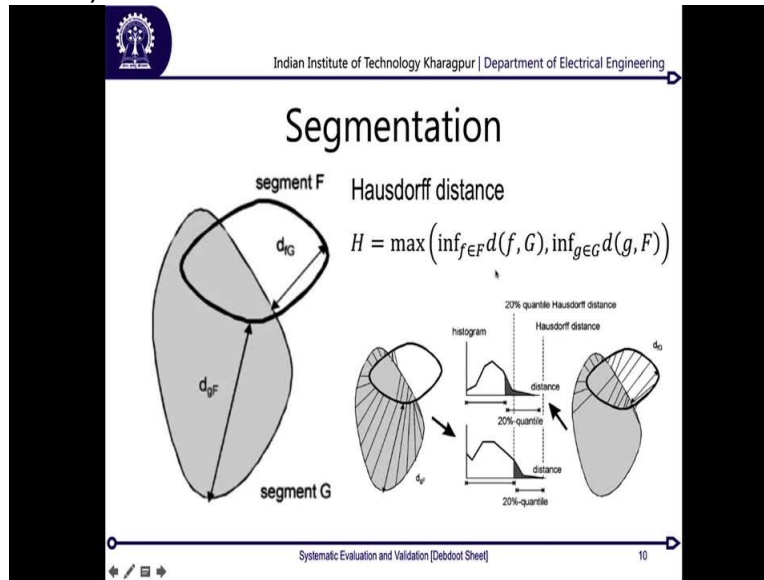
call this as g and x is the coordinate location. So the value of $g(x)$ is wherever the object is present is 1 and everywhere else it is 0.

And this set of all pixels which belong to this object over here is also called as A . Similarly there is a performance from your algorithm or whatever you are designing and that is called as the delineated object say and we represent in a similar way $f(x)$ and the set of these object points marked over there is called as B . Now in order to evaluate segmentation we have 4 different measures, one of them is called as Oversegmentation and this is basically the number of pixels which are segmented which are said to be belonging to the object as per your segmentation, but actually do not belong to the object.

So that is where I do an intersection with the compliment of this ground truth okay. You are always going to divide it by the number of samples present in the ground truth itself because that is the normalizing factor in all cases. Undersegmentation is the reverse of it which is it means that the pixels belong to the object as per the ground truth, but as per your segmentation they do not belong to the object, so you can have an undersegmentation performance over there. Your algorithm can have both over segmentation and undersegmentations.

the next one is called as a Dice coefficient and this is basically twice the intersection of the areas divided by the sum of the areas over there and the other one is called as a Jaccard coefficient which is which looks very similar to a dice coefficient except for the fact that this is an intersection divided by the union over there. So if you have a perfect segmentation this would end up being as 1. Now from there, these were all area based segmentation evaluation methods.

(Refer Slide Time: 32:14)



We come down to another one which is called as a distance based or when you have contours coming down, then how do you map down whether my contours are very closely located to each other or they are very far off. So to give you a very thing is that maybe your object is actually oval in shape and you found out a flowery like pattern around that oval. So all of your area wise overlap measures would give you a very high segmentation score but your boundary was not that smooth which you wanted to find.

Now for this you use this kind of a matrix which is called as a Hausdorff distance, now for Hausdorff distance what it does is quite simple so for every point you find out (eve) every point on your segmented region and every point on your actual ground truth G you would be find out which is the corresponding point over there. So you find out the largest distance of the corresponding points and then you are just going to take that maximum of the distance and this is the maximum error in terms of contour deviation which you would be plotting down.

Now over here it might appear to be very complicated but the idea is pretty simple. You just need to find out the centroids for both the objects and either in a clockwise direction or in a anti clockwise direction you just need to do the polar plot over there. Now once you find out the polar plot over there you have a point to point correspondences between the points coming down, just measure those units over there and after that you need to find out what is the maximum of that and that is a clear cut way of doing it.

(Refer Slide Time: 33:38)

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Classification

		Prediction	
		P	N
Ground Truth	P	TP	FN
	N	FP	TN

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- $Sensitivity = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{TN+FP}$
- $Precision = \frac{TP}{TP+FP}$
- $F - score = \frac{2TP}{2TP+FP+FN}$

Systematic Evaluation and Validation [Deebot Sheet] 11

From there we come into classification measures and so all of your problems are not exactly segmentations. Somewhere you might be giving an image as the input and you just want to find out what is the classification result over there. Now for typically a two class prediction model you will have a ground truth on which you have the positive class which is maybe the disease class and the negative class which are all the healthy class.

Now you would be predicting also as positive and negative. Our some of the samples which are actually positive you would also be predicting them as positive and they are called as true positives. Some samples which are actually negative you would also be predicting them as negative and that is called as true negative. There would be some samples which are actually negative but you are predicting them falsely as positives and they are called as false positives and there are some samples which are actually positive but you are falsely predicting them as negative and they are called as false negatives.

So this kind of a matrix is called as a confusion matrix for a two class problem. Now if you have a multiclass problem say three class problem then what you would end up getting is that you can take class 1 as your positive class ones and get a confusion matrix class 2 as your positive class ones get the confusion matrix. So class 1 can be misclassified as class 2 or class 3 so that is a negative problem over there. Class if you are taking class 2 as your positive class then the errors may be class 1 and class 2 which are class 1 and class 3 which are the erroneous results.

So similarly you will get down basically three different confusion matrixes over there coming down. Now for each of them you can find out the accuracy, sensitivity, specificity, precision and F one score according to the formulas which are listed over here. Now accuracy is an overall prediction which is how much you are going to predict both the positive class and the negative class in a good way, but that is not always true.

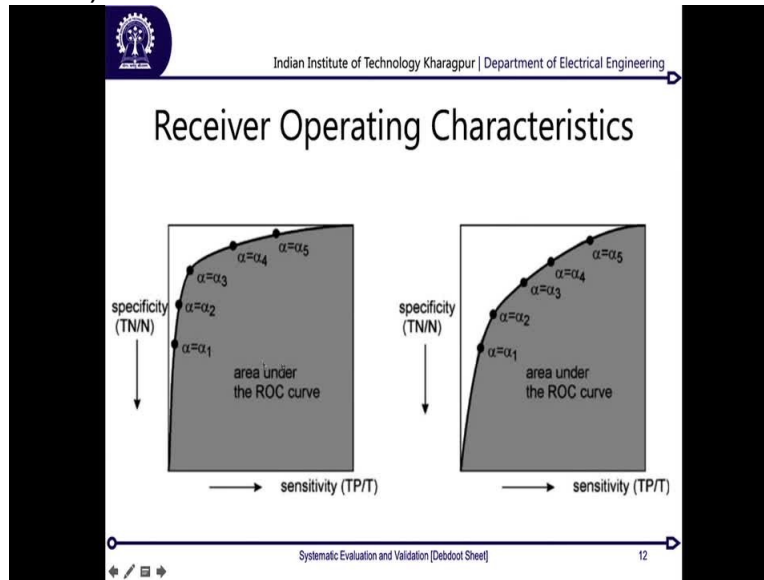
In some of the cases you would be looking more interested towards the positive class. I want to be very accurate for positive class, for disease scenarios obviously this rate is supposed to be close to 100 which means that my number of false negatives should actually be going down towards 0. Which means if there is a patient how has a disease I cannot tell that particular person that you are healthy to let him go away because that is where there will no check at any further point of time.

The other one is about specificity where you can have a lower score even the fact that what would happen is that you can call a lot of people as false positives which is a patient is does not have a disease but has been said to have a disease. Now in anyways after the automated process at any point of time the patient is obviously going down to a doctor who is an human expert and that human expert is going to find out that point of time.

So there are chances of the cost associated with doing a error over here is much lower but the cost associated with saying a diseased person that you are healthy is much higher because the person is going to die over here at least that thing is not going to happen over. So from there so these are about sensitivity, specificity and precision. So the final thing which a lot of people are liking this days is about F score because it is sort of a harmonic mean between sensitivity and precision.

So what this does is that you have a weightage towards the true positive places but you also take in to account the false positive and the false negatives and this is comes down to achieving sort of a regularized score when you have a imbalance data present in your classes as well. Because although you are training with a balanced biased and totally balanced data by doing augmentation on your data space but then you do not have balance data of during your test, so that is where you need to take care of it.

(Refer Slide Time: 37:20)



So from there we come down to this final thing which is called as Receiver Operating Characteristics. Now what it does is that you would look at the two axis one is the sensitivity, another is a specificity axis and what typically we wanted to do in this particular curve was that the idea was that say you have a probability coming down and you can take a decision by thresholding the probability at a point of time.

Now as you select this threshold say from point 1 till point 9 you would be getting down difference performances coming. Now you need to choose like which is my ideal performance coming down over there. Now if you are not able to understand as to which one to choose, a good way is basically trying to draw this sort of a plot and finding out the area under this one, now higher the area the better is the resilience to wherever you are trying to thresholding.

So what would ideally happen is that if ideal case would be a plot which is somewhat like this and that would mean that any point I choose to threshold I would actually be getting an area which is close to 1 and that means that my predictions are independent performance of my prediction is independent where I am thresholding and I am no more dependent on that one. So which classifies algorithms as being better or worst amongst each other.

There is another benefit of using this one as well. If you are trying to choose an ideal threshold then what you need to do is draw another 45 degree line over here and whenever it crosses this

particular line, you need to find out what is the possible value of threshold over there and just use that threshold and that is going to give you the best F score coming up at all point of time.

(Refer Slide Time: 38:51)

Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Folded Cross-validation

- Folding
 - Creating non-overlapping sample sets
 - Class 0 – N_0 Samples
 - Class 1 – N_1 Samples
 - Folds – k
 - Training samples per fold = $\frac{N_0 + N_1}{k} (k - 1)$
- Divide samples in k number of bags
 - k -th bag will contain
 - $\frac{N_0}{k}$ samples of Class 0
 - $\frac{N_1}{k}$ samples of Class 1
 - No bags will overlap

Systematic Evaluation and Validation (Debdoot Sheet) 13

Now once you know all the matrix and everything, the last but not the least is how are you going to tell about all unknown scenarios and the best way of doing it is actually to do a folded cross validation. Now how we do this is quite simple, so you are going to create non overlapping sample sets. So say I have a two class problems, class 0 and class 1 and for class 0 I have some N_0 number of samples and class 1 I have N_1 number of samples and I say that I am going to use k number of folds.

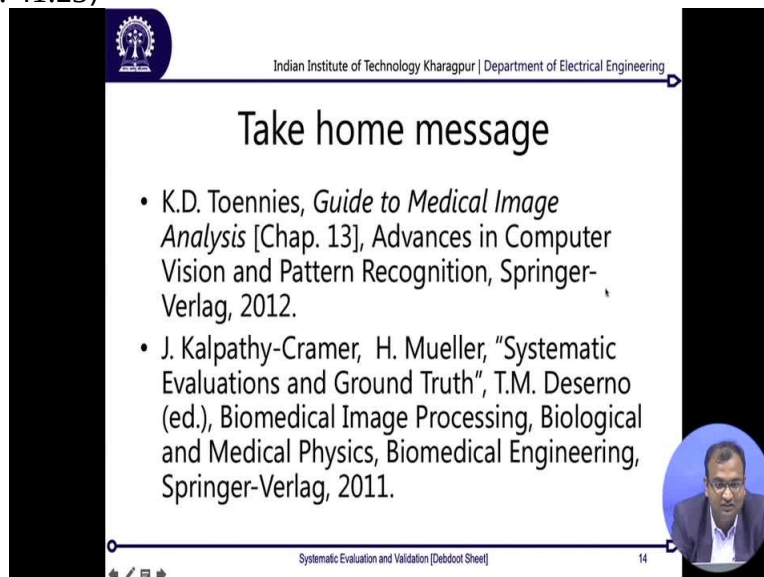
Now, if I am doing this one what I would do is that I am going to train with N_0 plus N_1 divided by k times k minus 1 number of fold number of samples and I am going to test on the N_0 plus N_1 by k number of samples. Now under a constraint that you will actually have to create k number of bags for doing the same, now what each bag means is that in each bag you will have N_0 by k number of samples for class 0 and N_1 by k number of samples for class 1 and also ensure that there are no samples which overlap between bags, which means two different bags are not going to consist the same number of samples, that is what is going to again introduce some sort of a variance some sort of a bias into system. So you cannot share between them.

Now what this is going to ensure is that at any point of time at any fold there will be a group of samples which have not been used for training the system at any point of time and you are going to test on them. So this is going to ensure that after doing a sort of a round robin method of testing so you start with fold one where you use set 1 in order to test and all the sets from 2 to k for training it. Then in fold 2 you are going to use set 2 for testing and for training you are going to use set 1 and set number 3 to k for training it.

So you are sort of doing a round robin over here and if you take a complete loop over there you would after some point of time find out that you have a sort of circle in which you are able to find out what is the total variance which your system might be having. So there would obviously be certain sets over there which will be biased towards a particular class and the other one towards another different class. So you will be seeing a tradeoff between the accuracies which you can achieve.

Now the best algorithm is that one which has the minimum amount of variance across all the folds and can give you the best performance. So this is what you are going to look at the total performance in the wild for your method.

(Refer Slide Time: 41:23)



Indian Institute of Technology Kharagpur | Department of Electrical Engineering

Take home message

- K.D. Toennies, *Guide to Medical Image Analysis* [Chap. 13], *Advances in Computer Vision and Pattern Recognition*, Springer-Verlag, 2012.
- J. Kalpathy-Cramer, H. Mueller, "Systematic Evaluations and Ground Truth", T.M. Deserno (ed.), *Biomedical Image Processing, Biological and Medical Physics*, Biomedical Engineering, Springer-Verlag, 2011.

Systematic Evaluation and Validation (Deebot Sheet) 14

From there the lecture ends and I basically come down to a take home message which I have so you can read for more details into this two texts from where most of the materials for making these slides were taken down and with that I come to an end and thank you.