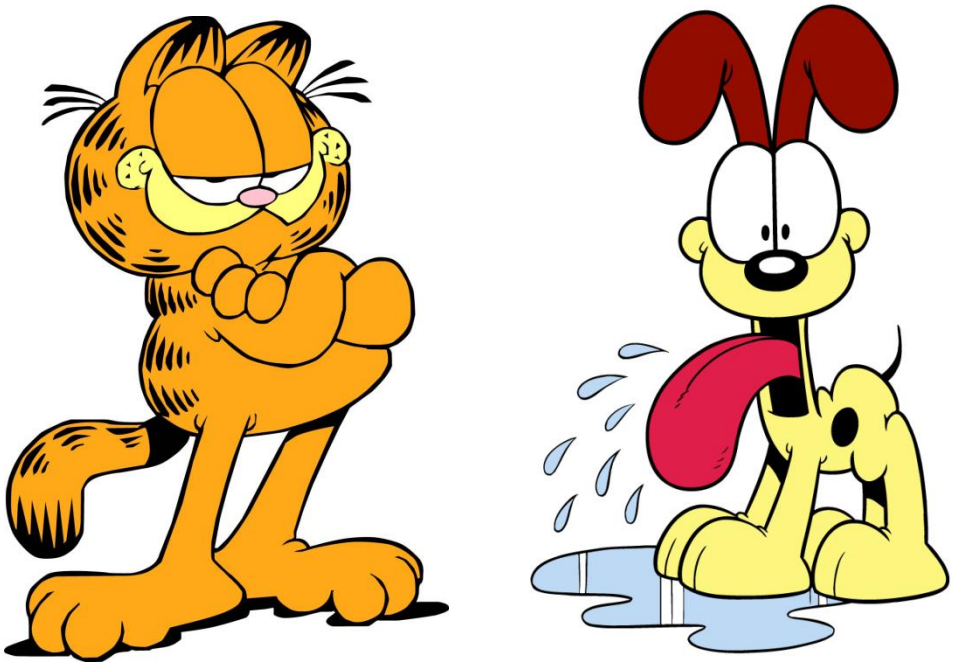
CLEMSON UNIVERSITY, SCHOOL OF COMPUTING

Data Mining Project

Can computer tell cat and dog apart?

Chao Huang, Di Zhang, Pei Pei

1/27/2014



Chao Huang: chaoh@g.clemson.edu

Pei Pei: peip@g.clemson.edu

Di Zhang: dzhang2@g.clemson.edu

1. the problem we want to work on

We want to write an algorithm using data mining technology to classify whether images contain either a dog or a cat. Our team chose this topic because the task that classify dogs or cat in images is easy for humans but not for computers. If we can figure this out, in the further, we can also develop this algorithm to classify some other things not only for dogs and cats. Therefore, we feel this task is very interesting, meaningful and challenge. Furthermore, if we complete it before the due day, our team plans to implement it as an APP (application). In this way, we can make this algorithm more visual which is also one of the reasons we chose this topic. About the datasets, we use include 25.000 images of dogs and cats. In these datasets, we also contain test data to train our algorithm. In the process of problem-solving, we should learn how to use Matlab to process image datasets.

2. the research approach including methodologies, data, evaluation methods

In this project, we do a project whose data and idea is provided by kaggle. Kaggle is a platform for predictive modeling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models. Thousands of data mining scientists work together in the same projects. So that, we can learn a lot from others.

The evaluation method contains three questions:

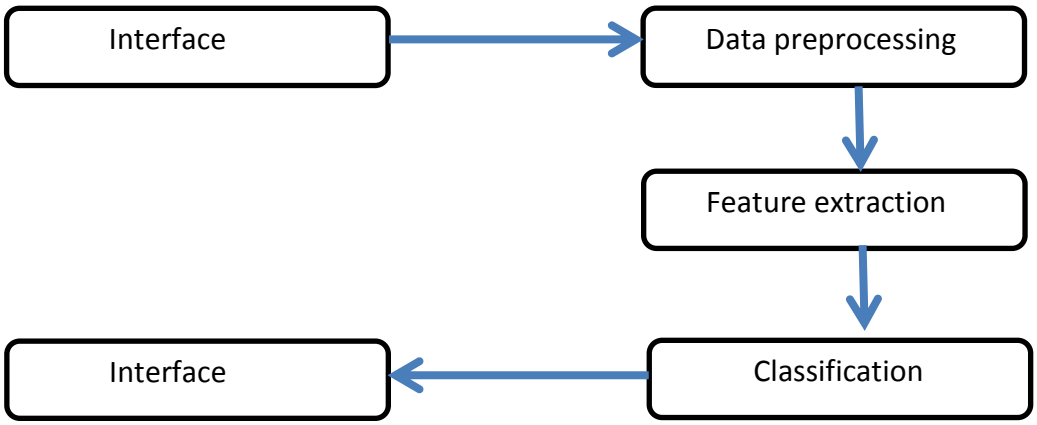
A. The classification result: Can we tell cat and dog a part correctly?
B. The time consuming: Can we give the answer as quick as we can?
C. User interface: Can we let the user use our methods any time they want?

We have to say, that the third question is an optional for us. It depends on the first two questions. If we can successfully finish the basic requirement, we will try to finish the third question.

Matlab and C++ (or Java) will be used in our project. Matlab will be used to test our algorithm correctness. C++ (or Java) will be used to construct the user's interface.

There are many mature algorithms that focus on image processing such as: image segmentation, such as hot, sift image processing, principle component analysis, and even there is a new algorithm named 'cat face' analysis. I think these basic methods will be helpful for our project. We will try some or all above algorithms and make a comparison among them.

The following is the procedure of our project:



3. ***the expected results, such as research goals, system to be developed***

In this project the quality of the result is determined by the accuracy of the algorithm, i.e., the percentage of correctly classified pictures among all the training data; The algorithm built at this moment to classify the pictures can reach a percentage of 80%, thus the goal of this project is to develop an faster and more efficient algorithm to reach a better result.

What we do expect is to add more categories for the algorithm to classify; By taking more visual factor into account and analysis more parameters synthetically we can get a more meaty result. This part of work will do according to the progress of the program.

4. ***Alternative approach if your proposed approach does not work as expected, i.e., your backup plan.***

   If we fail to achieve the object we have make for ourselves. We will give the result of each sub-procedure such as the 'Eigenface' for each picture. At the time, we will give an analysis for each method about why it does not work well for this job.

5. ***project milestones***

time	work
01/14—01/26	making team, determining topics and then writing a proposal
01/26---03/01	Data collection and clean dataset
beginning of March	intermediate project report
March--- April 25	final report

6. ***roles of team members in your project***

Member	work
Chao Huang	Algorithm design, implementation and basic interface design
Di Zhang	User interface instruction and written materials
Pei Pei	Algorithm design, implementation and written materials