Classificação Hierárquica de Caracteres Matemáticos

Ricardo Alexandre Bastos[2] *
Orientadora: Profa. Dra. Nina Sumiko Tomita Hirata
Instituto de Matemática e Estatística
Universidade de São Paulo

9 de Fevereiro de 2010

1 Parte Subjetiva

Essa seção contém os relatos sobre a execução e desenvolvimento do trabalho e das melhorias feita no projeto *Math-Picasso* e sua conversão para a estrutura do *Express-Math*.

Gostaria de aproveitar também para agradecer a Professora Nina Hirata, que se dispôs a fazer a orientação do projeto.

1.1 Desafios e Desenvolvimento

Durante o início do ano ocorreram problemas com minhas idéias iniciais para o projeto, então comecei a procurar um novo projeto. Então consegui conversar com a professora Nina e entrar no projeto *Math-Express*, mais precisamente na parte de reconhecimento de escrita. Então comecei o desenvolvimento do projeto com o aluno Breno.

O começo do projeto foi bem lento e demorado: como utilizamos o código da interface do *Math-Picasso*, foi necessário compreender o código do projeto. Que era mal documentado e com um código pouco claro. Foi uma atividade maçante fazer a refatoração do código: tivemos que eliminar diversas interfaces inúteis e alterar a interface gráfica. Essa limpeza do projeto

*1 rabastos44@gmail.com

inicial fez com que adiássemos o desenvolvimento do reconhecedor de caracteres.

Em paralelo a essa limpeza do código, comecei a pesquisar algumas características utilizadas em reconhecimento de caracteres. A pesquisa foi interessante e produtiva. Depois de um certo tempo, notamos que diversas características eram bastante frequentes em outros trabalhos, então escolhemos algumas bastante utilizadas e desenvolvemos algumas que pareciam promissoras em alguns casos. Como escolhemos características que definiam os caracteres para nós, foi necessário descobrir uma maneira de extraí-las do conjunto de pontos, a explicação óbvia foi utilizar uma representação de traços.

Depois da implementação das características, comecei a pesquisa por um novo classificador. Como o objetivo era utilizar alguma abordagem que utilizasse alguma forma de hierarquia para diminuir o processamento, depois de algumas pesquisas decidimos por implementar o BHC.

A implementação do BHC foi bastante desafiadora. O artigo que o descreve é bem detalhado mas a adaptação do algoritmo genérico para um especifico para caracteres levou algumas semanas. Um dos problemas da implementação na verdade foi a implementação das dependências do BHC, o discriminante de Fisher e da função de densidade de probabilidade de uma curva Normal. Como as funções mencionadas utilizam manipulação de matrizes, decidi encontrar uma biblioteca que conseguisse manipular matrizes de forma eficiente. Encontrei a biblioteca Math da *Commons Apache* que funcionou de maneira realmente boa, até que foi necessário extrair os autovetores e autovalores de uma matriz não-simétrica - funcionalidade não implementada na Math. Depois de um tempo procurando, encontrei a biblioteca *JAMA* (Java Matrix). E tive que alterar a implementação para a biblioteca *JAMA*.

Depois da implementação pronta, faltava testar a eficiência do algoritmo. Infelizmente não possuía o tipo mais natural de entrada, um *tablet*, para capturar a escrita. Então fiz testes com um mouse e com testes não coletados por mim, até conseguirmos um *tablet* para teste e refinamento do algoritmo.

A ultima parte que faltava era integrar o BHC no Express-Math, o que foi feito com grande facilidade dada a refatoração anterior. Infelizmente não foi possível utilizar o novo reconhecedor com a segunda parte de projeto - o reconhecimento de estrutura - pois não conseguimos um treinamento eficiente de todas os caracteres necessários.

1.2 Matérias Relevantes no Trabalho

Muitas das matérias cursadas previamente que foram utilizadas no projeto estão listadas abaixo:

Mac0110 e Mac0122 foram muito úteis, por serem matérias introdutórias e por meu primeiro contato com Java ter sido em Mac0110.

Mae0121 e Mae0212 as matérias introdutórias de estatística foram muito úteis. Dada a natureza do projeto de aprendizagem computacional, a parte de estatística foi realmente importante tanto na modelagem da árvore de classes e no classificador.

Mac0211 e Mac0242 muitas das ferramentas utilizadas para o desenvolvimento do projeto aprendi nas matérias de laboratório.

Mac0323 os conhecimentos em estruturas de dados definiram a maneira de organizar a árvore como um heap e manipular estruturas em árvores.

Mat0139 a noção de espaços vetoriais e projeção utilizada durante a seleção de características, além das decomposições utilizadas para acelerar o algoritmo, no cálculo de inversões e no cálculo do discriminante de Fisher.

Mac0300 os algoritmos utilizados para fazer a decomposição em autovetores e autovalores, mesmo que eu não os tenha implementados, foram conhecidos em Métodos Numéricos de Álgebra Linear

Mac460 Aprendizagem Computacional: muitos dos conhecimentos necessários para a implementação, teste e utilização de um classificador foram vistos nesta matéria.

1.3 Trabalhos Futuros

A área de aprendizagem de maquina é muito interessante e possui muitas utilidades. Como eu só trabalhei com um pequeno grupo das possibilidades dentro da área, teria que continuar estudando outros algoritmos, alguns citados no trabalho mas vistos sem aprofundamento. Além disso, talvez escolher conhecer as diversas aplicações.

É uma área que ainda precisa de pesquisas e onde não há respostas prontas, portanto, para tentar avançar nela é necessário um conhecimento profundo dentro dela.