ELABORAÇÃO DE BASE DE DADOS PARA FINS DE PESQUISA EM CLASSIFICAÇÃO DE ÁUDIO

Viviane Shiraishi Besson (PIBIC/CNPq/FA/Uem), Yandre Maldonado e Gomes da Costa (Orientador), e-mail: ra93298@uem.br.

Universidade Estadual de Maringá / Centro de Tecnologia / Maringá, PR.

Ciências Exatas e da Terra / Ciência da Computação

Palavras-chave: base de áudio, classificação de áudio, anotação de áudio.

Resumo:

A geração de recursos para apoio ao desenvolvimento de pesquisas em reconhecimento de padrões é de grande importância para viabilizar novas descobertas neste contexto de investigação científica. Um dos empecilhos mais conhecidos por pesquisadores da área é, em muitos casos, a dificuldade para se obter uma base de dados com características apropriadas para o desenvolvimento de trabalhos de pesquisa. Este projeto teve como objetivo colaborar com a criação da *Freesound Datasets* e desenvolver alguns experimentos iniciais para a classificação de amostras de áudio a partir de subconjuntos de dados contidos nela.

Introdução

Nos tempos atuais é muito grande a quantidade de gravações de áudio geradas diariamente em função da popularização de dispositivos de captura, como *smartphones*. Além disso, a disseminação do uso de ferramentas de comunicação e compartilhamento, como *whatsapp* e redes sociais, favorece a circulação e o acesso a esses conteúdos. Esse cenário traz consigo inúmeras possibilidades de investigação científica, no sentido de propor novos métodos que permitam viabilizar aplicações para o grande e crescente número de usuários desses recursos.

A identificação do ambiente ou contexto em que determinada amostra de áudio foi gravada é uma tarefa que vem despertando grande interesse na comunidade de pesquisa em classificação de áudio, com aplicações que vão desde a identificação automática de conteúdo multimídia, até aplicações para cuidados em saúde sensíveis ao contexto.

Em 2017, o *Music Technology Group* da *Universitat Pompeu Fabra* de Barcelona lançou o projeto *Freesound Datasets* (FONSECA et al, 2017). Em um primeiro momento, os pesquisadores recrutam voluntários para participarem da anotação dos dados disponibilizados por colaboradores.

Este projeto de pesquisa tem como objetivo a participação na elaboração da *Freesound Datasets* e a familiarização com o protocolo básico de anotação de itens para a criação de grandes bases de dados. Adicionalmente, desenvolveu-se classificadores de áudio automáticos no cenário multirrótulo a fim de gerar resultados de base que sejam úteis para o desenvolvimento de pesquisas futuras.

Materiais e métodos

Participação na elaboração da Freesound Datasets

A participação na elaboração do *Freesound Dataset* constitui-se na classificação de áudios rotulados de acordo com cerca de 50 categorias diferentes. Para cada categoria, é disponível uma breve explicação do som, a árvore hierárquica, categorias-irmãs, exemplos de som, quantidade de amostra e anotações validadas por voluntários. Para cada anotação, são disponibilizadas em torno de 72 amostras de áudio que poderiam ser associadas àquela categoria. Com isso, para cada amostra era dado as alternativas: presente e predominante (a amostra contém áudio referente à categoria e é predominante), presente mas não predominante (a amostra contém áudio referente à categoria, porém contém outros sons não pertencentes à categoria, como som de fundo), não presente (amostra não contém som da categoria) e incerto (não há a certeza se a amostra pertence ou não a categoria). Ou seja, verificou-se se as anotações geradas realmente pertenciam à categoria.
Com isso, das 50 categorias classificadas, obteve-se no total 1571 amostras em áudio e 34 rótulos, sendo eles as categorias pertencentes à som de animais (pássaro, voo de pássaros/bater de asas, som do galopar de cavalos, cacarejar de galo, rugido, mugido), som humano (crianças cantando, crianças brincando, fala, voz humana, gargarejo, risada, ronco), som natural (fogo, água natural, trovão), som de objetos (motosserra, som de secador de cabelo, som ao andar de skate, moeda caindo, corrida de carro, água de torneira), música (bateria (instrumento), acorde, caixa da bateria, música triste, música feliz, piano, instrumento musical, nota musical), som de fundo/ambiente (repercussão, vibração, som dentro de um cômodo grande ou salão, som em espaço público). As amostras variam de 1 segundo até 90 segundos, a qual mais de 50% das amostras possui duração de até 10 segundos. Além disso, a distribuição das amostras entre as categorias é irregular, ou seja, alguns rótulos possuem poucas instâncias. O conjunto de dados criado foi chamado de *Multi-Label Acoustic Scene*[1] e está disponível para a comunidade de pesquisa para o desenvolvimento de trabalhos futuros.

Experimentos

Primeiramente, converteu-se as amostras de áudios obtidas para a classificação de *Freesound Datasets* para o formato *wave* e áudio mono utilizando o programa *SoundConverter* e extraiu-se espectrogramas monocromáticos com o uso do *software Sound eXchange* (SoX) com taxa de amostragem de 14 kHz. Para a

[1] https://sites.google.com/view/mulasc/home

extração de vetores de características dos espectrogramas gerados, utilizou-se o descritor *Local Binary Pattern* (LBP) com oito vizinhos e distância dois.
A classificação das amostras foi realizada com a utilização do *framework* MULAN (TSOUMAKAS et al, 2010) que é uma biblioteca Java para conjunto de dados multirrótulos. Para isso, criou-se dois arquivos necessários: um no formato *arff* contendo os vetores de descrição de características e a qual rótulo cada amostra pertence (0 para a ausência do rótulo na amostra e 1 para a presença) e outro no formato *XML* contendo os rótulos e a hierarquia (caso exista) entre eles. Os algoritmos executados foram *Binary Relevance* (BR), *Multi-Label k-Nearest Neighbors* (MLkNN) para 8, 9, 10, 11 e 12 vizinhos e *Instance-based logistic regression for multi-label classification* (IBLR-ML E IBLR-ML+) disponibilizados no *framework* e modificados para realizar o método validação cruzada de 4-fold.

Resultados e Discussão

A base criada e utilizada neste projeto é do domínio de áudio com 1571 amostras e 34 rótulos. Outras propriedades que caracterizam a base estão sumarizadas na Tabela 1.

Tabela 1 – Indicadores da base multirrótulos.

Cardinalidade	Densidade	Diversidade	Proporção de diversidade
2,409	0,071	206	0,131

Os resultados obtidos nos experimentos foram coletados e dispostos na Tabela 2. Utilizou-se métricas baseada em ranking (precisão média e perda de Hamming) e baseada em rótulo (*micro-averaged precision* e *microAUC*).

Tabela 2 – Métricas para cada algoritmo classificador.

Algoritmos classificadores		Precisão média	*Micro-averaged precision*	Perda de Hamming	*MicroAUC*
BR		0,3298	0,1445	0,3164	0,7676
MLkNN	8 vizinhos	0,6248	0,7319	0,0538	0,9131
	9 vizinhos	0,6256	0,7188	0,0543	0,9134
	10 vizinhos	0,6292	0,7263	0,0539	0,9136
	11 vizinhos	0,6207	0,7313	0,0542	0,9139
	12 vizinhos	0,6206	0,7265	0,0547	0,9146
IBLR-ML		0,5907	0,5780	0,0640	0,8970
IBLR-ML+		0,5701	0,5006	0,0708	0,8863

Pode-se observar que o algoritmo MLkNN possui os melhores valores de precisão média e *micro-averaged precision* – precisão média em todos os pares amostras/rótulos –, 62,92% para 10 vizinhos e 73,19% para 8 vizinhos respectivamente. Em contrapartida, o algoritmo BR teve o pior desempenho, com

precisão média de 32,98%. Para os valores de *MicroAUC*, o classificador MLkNN possui a melhor performance entre eles com valores acima de 90% para 8-12 vizinhos. Além disso, o algoritmo MLkNN com 8 e 10 vizinhos possui a menor perda de Hamming, 5,38% e 5,39% respectivamente. Ou seja, teve a menor porcentagem de rótulos previstos incorretamente em relação ao total de rótulos. Enquanto, novamente BR teve a pior performance com a maior taxa de 31,64%.
Com os valores de *micro-averaged precision* mostra que a quantidade desigual de amostras em cada classe, como por exemplo, gargarejo e rugido que possuem somente duas amostras em cada, influencia na precisão, uma vez que não há a possibilidade de distribuição uniforme entre as quatro pastas no método de validação cruzada. Além disso, o baixo desempenho do classificador BR é que ele constrói um classificador binário para cada rótulo independentemente, podendo não considerar correlações entre os rótulos.

Conclusões

Este projeto teve como objetivo a participação na construção do *Freesound Datasets*, além da elaboração de experimentos nas amostras adquiridas. Para isso, verificou-se anotações geradas para diversas categorias e obteve-se amostras, as quais extraiu-se espectrogramas, vetores de características e utilizou-se *framework* MULAN para as classificações. Entre os algoritmos classificadores executados, MLkNN, principalmente com 10 vizinhos, teve a melhor performance. Enquanto *Binary Relevance*, por tratar individualmente cada rótulo, teve um pior resultado. Além disso, a desigual quantidade de amostras entre as categorias e, em alguns casos, poucas amostras, influenciou em uma baixa precisão.

Agradecimentos

Agradeço ao professor Yandre Maldonado e Gomes da Costa e à Universidade Estadual de Maringá pela oportunidade. E à Fundação Araucária pelo suporte financeiro.

Referências

Fonseca, E., Pons J., Favory X., Font F., Bogdanov D., Ferraro A., Oramas S., Porter A., & Serra X. **Freesound Datasets: A Platform for the Creation of Open Audio Datasets**. 18th International Society for Music Information Retrieval Conference, 2017.

Tsoumakas, G., Katakis, I., Vlahavas, I. Mining Multi-label Data. In **Data Mining and Knowledge Discovery Handbook**. Springer, 2010.