34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Multi-Label Classification for the Analysis of Human Motion Quality

Portia E. Taylor, Gustavo J.M. Almeida, Jessica K. Hodgins, and Takeo Kanade

Abstract— **Knowing how well an activity is performed is important for home rehabilitation. We would like to not only know if a motion is being performed correctly, but also in what way the motion is incorrect so that we may provide feedback to the user. This paper describes methods for assessing human motion quality using body-worn tri-axial accelerometers and gyroscopes. We use multi-label classifiers to detect subtle errors in exercise performances of eight individuals with knee osteoarthritis, a degenerative disease of the cartilage. We present results obtained using various machine learning methods with decision tree base classifiers. The classifier can detect classes in multi-label data with 75% sensitivity, 90% specificity and 80% accuracy. The methods presented here form the basis for an at-home rehabilitation device that will recognize errors in patient exercise performance, provide appropriate feedback on the performance, and motivate the patient to continue the prescribed regimen.**

I. Introduction

Approximately 27 million people in the United States have been diagnosed with osteoarthritis (OA) [11]. Knee OA is the most common form of OA and is a degenerative disease of the cartilage of the knee. There is no cure for knee OA and the exact cause is unknown; however, obesity, prior injury, and aging have all been identified as factors contributing to the disease. Knee osteoarthritis is often associated with aging, with 33% of patients being over the age of 63 [1].

Patients diagnosed with knee osteoarthritis are often prescribed therapeutic exercises to be completed in the home. Research has shown exercise to be an effective treatment for increasing joint mobility and decreasing pain and stiffness [8]. Journals or exercise logs are sometimes completed by patients for monitoring adherence to the exercise program and also to report self-perceived effectiveness of the exercises. Often patients fail to adhere to the prescribed program or perform the exercises incorrectly. To our knowledge, there is no system currently available that provides a quantitative measure of the quality of human motion achieved while performing therapeutic home exercise.

This paper is organized as follows. We first discuss related work in Section II. Section III describes the details of our experimental set up, data collection procedure, and learning methods used. In Section V, we present the results obtained from our system and discuss our findings. We conclude in Section VI with some limitations of our system and a discussion of future work.

This work was supported by the Quality of Life Technology Center under National Science Foundation award #0540865 and National Science Foundation award #0931999

P.E. Taylor is with the Biomedical Engineering Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA pet@cs.cmu.edu

G.J.M. Almedia is with the Department of Physical Therapy, University of Pittsburgh, Pittsburgh, PA 15213, USA gja4@pitt.edu

J.K. Hodgins is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA jkh@cs.cmu.edu

T. Kanade is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA tk@cs.cmu.edu

II. Related Work

A. Sensor-Based Rehabilitation Systems

Sensor-based systems for rehabilitation in the home are increasingly mentioned in the literature. Jovanov and colleagues describe a platform for a computer assisted rehabilitation system using off-the-shelf motion and biological sensors [4]. The system would collect information about a user's rehabilitation progress, provide feedback to the user, and provide recorded data to medical servers. Tseng and colleagues also describe the structure of a home rehabilitation system based on accelerometers and compasses [13]. The system provides instruction to the user on exercise in a game-like fashion. Tseng was also interested in evaluating the quality of motion performed during exercise; however, there has not yet been any published work using the system on patients.

Melzi and colleagues developed a virtual training system utilizing wireless two-axis accelerometers to capture human movement [6]. The researchers were able to extract the number of completed repetitions, speed of performance, and fluidity of movement during the performance of a biceps curl. Feedback was provided to the user via a colored indicator bar and video of a trainer's performance.

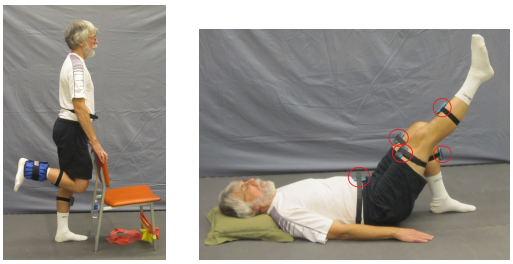
Taylor and colleagues demonstrated that tri-axial accelerometers and simple machine learning algorithms could be used for measuring human motion quality by conducting experiments on healthy college students performing exercises for knee OA [12]. They collected data using five sensors and utilized the AdaBoost classifier. They obtain high rates in specificity and sensitivity; however, their approach transformed the multi-label data set into one-versus-all single label construct rather than performing a true multi-label classification as we do here.

B. Multi-Label Classification

In traditional classification problems, each training example $x \in X$ is associated with a single label $\lambda \in L$. We instead take a multi-label approach to the analysis of human motion quality in therapeutic home exercise. Our problem requires a multi-label classifier because a patient may commit one or more errors simultaneously in a single performance (repetition) of an exercise. All errors should be detected although some may not require feedback.

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

Fig. 1: Wireless Sensor Node



(a) SHC (b) SLR

Fig. 2: The standing hamstring curl (SHC) and straight leg raise (SLR) exercises for knee OA

There is a large body of work on multi-label learning for text and web-page classification. Tsoumakas and colleagues review various multi-label learning techniques [14]. Schapire and colleagues introduced variants of their popular AdaBoost algorithm to support multi-label classification [10]. One of the most widely used is the Binary Relevance (BR) transformation. In BR, the multi-labeled data set is transformed into q data sets, one for each label in L. The transformed data have all the instances of the original data set with a positive label if the instance contains the current label and a negative label otherwise. BR is a one-versus-all classification technique. A new example is classified by taking the union of all labels predicted by each of the q classifiers.

Binary Relevance is not capable of representing dependencies between labels. To combat this problem, Read and colleagues developed the Classifier Chain (CC) model [9]. The CC model utilizes the low computational complexity of the BR classifier while adding a mechanism for retaining label associations within the data. In CC, q binary classifiers are built and linked in a chain $C_1, \ldots, C_q$ where each classifier C_j deals with the binary problem associated with label $\lambda_j \in L$, sharing the label associations learned in the previous chain links $\lambda_1, \ldots, \lambda_{(j-1)}$. To classify a new example, the process begins at C_1 and determines $P(\lambda_1|x)$, the probability of label λ_1 given example x. This process continues down the chain for every $C_2, \ldots, C_q$ predicting $P(\lambda_j|x_i, \lambda_1, \ldots, \lambda_{(j-1)})$. Information about previous binary associations are passed down the chain thus taking into account any label correlations. The Ensembles of Classifier Chains (EEC) method trains r classifier chain models, each with a random ordering of classifier chains and a subset of the training data.

A second method for handling potential correlations between labels uses the Label Powerset method [16]. From a set of labels L, the Label Powerset (LP) method treats each subset (or labelset) of L as a unique class in a single-labeled classification problem. The advantage of this approach is that label correlations can be represented within the individual labelsets. The RAkEL, Random k-Labelsets, method takes a set of L labels and randomly breaks it into many smaller labelsets. A LP classifier is then trained for each of the resulting labelsets and when a new example is observed, decisions are made by combining the LP models.

III. Experimental Setup

A. Hardware and Software

Multiple sensing nodes were used for data collection. Each node contains a tri-axial accelerometer (ADXL335 from Analog Devices) and a tri-axial gyroscope (Two-axis LPR530AL and one-axis LY530ALH from ST Microelectronics) for a total of six degrees of motion tracking. The module used is the ArduIMU+ V2 from SparkFun Electronics and features an Arduino-compatible Atmega328 at 16mhz processor. The sensors have an acceleration range of $\pm$ 4.5 g and an angular velocity range of $\pm$ 300 degrees per second. The sensor node is 40 x 30 x 20 mm and weighs 21 grams (including battery). Wireless communication between each node and the computer is performed with the 2.4 GHz XBee 1 mW Series 1 module by Digi. Figure 1 shows the nodes used in this work.

B. Data Collection

Data was collected from eight individuals, eleven female and four male (mean age 73.5$\pm$10.3), with clinically diagnosed knee osteoarthritis. Participants wore sensors placed on the thigh and shin of both legs mid-way between the joints and the front center of the waist (five sensors total). The attachment location for each sensor is shown in Figure 2(b). In this paper, we study two exercises commonly prescribed to people with knee osteoarthritis (OA): the standing hamstring curl (SHC) and straight leg raise (SLR) (Figure 2). For the SHC, participants wore a cuff weight of three to five pounds around the ankle of the working leg. Each participant was instructed on the correct form of each exercise by a physical therapist. Participants then completed three sets of ten repetitions of each exercise, first on the right leg then on the left. Some participants were unable to complete three sets of certain exercises. We have a total of 460 repetitions for the SHC and 440 repetitions for the SLR. We also used exercise data from Taylor and colleagues [12]. This data was collected from healthy college students completing the SHC and SLR exercises. The study described in this paper has been approved under CMU IRB protocol HS11-746.

C. Feature Selection and Data Labeling

Features were derived individually for each component of the five accelerometers and gyroscopes on a per-sensor

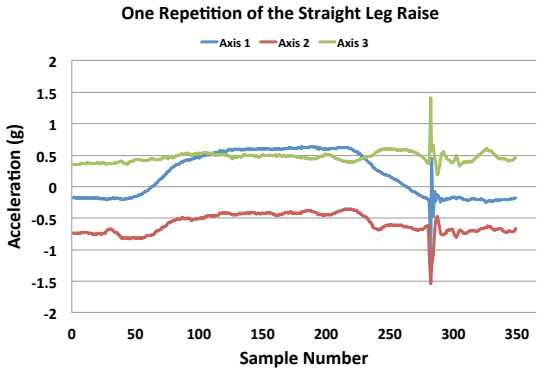


Fig. 3: Accelerometer data from Subject 8 SLR

basis for each repetition. We computed mean acceleration, mean angular velocity, the total number of examples in each repetition, and the first five frequency components for each axis of sensing. Preece and colleagues [7] provide an overview of the different feature sets used in the activity recognition literature and our features are among the ones reported there.

Data was automatically segmented into repetitions. Data from the sensor located on the shin of the working leg was smoothed using the logically weighted polynomial regression (loess) method with a second order polynomial. The index of the peak of each repetition was found. The beginning and end points of each repetition were determined by taking the indices of the minimum values (to the left and right of the peak) with a minimum segmentation of one second. Figure 3 shows a repetition resulting from our segmentation method. The is higher acceleration at the end of the repetition corresponds to an error where this patient dropped the foot to the floor at the end of the SLR exercise.

To provide ground truth labels, a physical therapist was given videos from a small subset of the collected data (7%). These videos represented performances from different subjects and exercises and were chosen based on the range of errors observed. We asked the expert to score (using a numeric scale) the labels for each exercise by occurrence in real-world observation and by severity of the error. A non-expert used the labels provided by and guidance from the expert to label the remaining data.

IV. Classification Methods

We use this data for the testing and training of multi-labeled classifiers that can identify multiple errors in the performance of an exercise. We train a separate classifier for each type of exercise.

A. Classification

We use the AdaBoost, Binary Relevance, Ensembles of Classifier Chains, and RA*k*EL methods for classification. AdaBoost was used as the classifier in single label experiments and as a base classifier for the BR and ECC methods in the multi-label experiments. Given a set of training data, the AdaBoost algorithm constructs a strong classifier by linearly combining various weak classifiers selected during the training process [2].

For the label-based experiments, we train a classifier for each label in the multi-label data using a one-versus-all approach. Data sets were created containing all repetitions from the original data set labeled as 1 if the example contained the label and 0 otherwise. Each of the data sets were trained separately using AdaBoost with a depth one decision tree base.

The multi-label classifiers were applied directly to the multi-label data set. For BR and Ensembles of ECC, boosting decision trees was used as a base classifier. For RA*k*EL, the LabelPowerset method was used with decision trees. The learners and base learners used were chosen for their intuitive nature, simple design, and relatively low computational costs.

AdaBoost and the C4.5 decision tree algorithm were both implemented using the WEKA software package [3]. The multi-label learners are implemented in the Mulan Java library [15].

B. Evaluation Metrics

In order to evaluate the classification performance of our methods on exercise data, we performed cross validation. The data set was randomized before being split into ten testing and training sets (folds). Each testing set contained data not present in the training set of that fold. For the label-based experiments, we report sensitivity and specificity resulting from the cross-validation procedure for each label. For the multi-label classifiers we report subset accuracy, hamming loss, and macro-sensitivity and specificity.

Subset Accuracy, or classification accuracy, is a strict metric that requires the predicted set to be an exact match to the actual set of labels. This accuracy is similar to the measure of accuracy used in binary classification.

Hamming Loss is a metric that takes into account how many times a label not belonging to an example is predicted or a label belonging to an example is not predicted [5]. Hamming loss is defined as

$$\frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \Delta Z_i|}{M} \tag{1}$$

where Y_i is the set of actual labels and Z_i is the set of predicted labels for the ith example. Perfect performance has a Hamming loss of zero.

Macro-averaged sensitivity and specificity are computed by taking the one-versus-all confusion matrices of each label and averaging them. The number of true positives, false positives, true negatives and false negatives for class λ is represented by TP_λ, FP_λ, TN_λ, and FN_λ respectively. The macro-averaged sensitivity, also known as recall, is defined as

$$\frac{1}{q}\sum_{\lambda=1}^{q}\frac{TP_\lambda}{TP_\lambda + FN_\lambda}, \tag{2}$$

where q is the number of classes. This metric give us an indication of how well our classifier is detecting the class of

Binary Relevance + AdaBoost		
	SHC	SLR
Hamming Loss	0.06	0.10
Subset Accuracy (%)	0.84	0.67
Example-Based Sensitivity (%)	0.93	0.88
Example-Based Specificity (%)	0.96	0.94
Macro-Sensitivity (%)	0.85	0.85
Macro-Specificity (%)	0.93	0.89

TABLE I: Multi-Label Results of Standing Hamstring Curl and Straight Leg Raise Data in [12]

interest that, in most cases, represents an error in an exercise performance. The macro-averaged specificity in equation (3) tells us how well the classifier can detect the negative class.

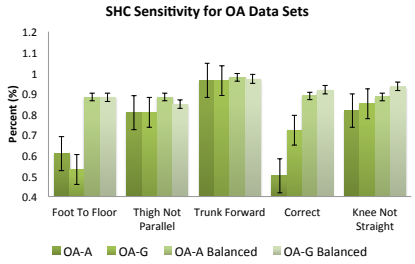
$$\frac{1}{q}\sum_{\lambda=1}^{q}\frac{TN_\lambda}{TN_\lambda + FP_\lambda} \tag{3}$$

V. Results

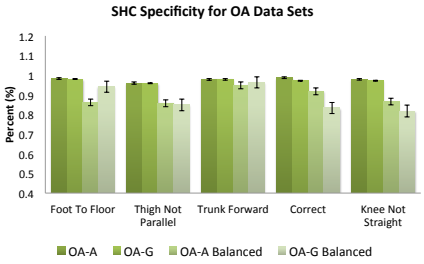
Shown in Table I are the results from using AdaBoost.M1 as a base classifier for the Binary Relevance algorithm using the healthy subject data set. This result is what could have been obtained in [12] if a Binary Relevance with AdaBoost.M1 base classifier had been used. For the remaining results, we use two different osteoarthritis data sets named OA-A (accelerometer only) and OA-G (accelerometer and gyroscope). Results are also presented for an equal number of positive and negative examples for a class (balanced).

Figures 4 and 5 shows the results from cross validation for each label on knee OA data of subjects performing the standing hamstring curl (SHC) and straight leg raise (SLR). An AdaBoost classifier was trained per label and the sensitivity and specificity were obtained. Sensitivity in Figure 4(a) is high when looking at classes thighNotParallel, trunkForward and kneeNotStraight. FootToFloor and Correct had the lowest sensitivity rates when using the OA-G set, with Correct having a higher rate of 72%. For the balanced dataset, the sensitivity rates increase for both OA-A and OA-G. TrunkForward has been identified by our physical therapist collaborator as being a commonly occurring error. We are consistently able to correctly identify that class with both OA-A and OA-G data in unbalanced and balanced form. In Figure 4(b), specificity for all classes of the SHC is above 95% for the unbalanced data. The balanced data averaged 89% for the OA-A data and 88% for the OA-G data. We observed lower rates for sensitivity and specificity in both the OA-A and OA-G balanced data sets. For each class, except FootToFloor of the SLR, the data contained more examples of the negative class. To balance the data, we generate a random, balanced subsampling of the data. This results in much smaller training sets (often only 30% of the dataset).

In Figure 5(a), sensitivity of the OA-A and OA-G data sets are high for most classes. Correct in the OA-A data was 70% and rose to 97% for OA-G. Similar increases were seen in KneeNotFullExtend and Overshooting. The number of examples for class FootToFloor account for 79% of the



(a) Sensitivity



(b) Specificity

Fig. 4: SHC results from 10-fold cross validation

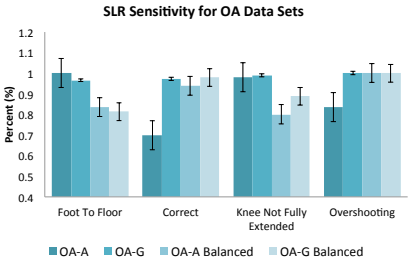
Osteoarthritis Data with Accelerometer and Gyro (OA-G)			
	BR	ECC	RAkEL
Hamming Loss	0.04	0.05	0.03
Subset Accuracy (%)	0.83	0.82	0.86
Macro-Sensitivity (%)	0.78	0.75	0.84
Macro-Specificity (%)	0.98	0.98	0.99

TABLE II: Multi-Label Results for the Standing Hamstring Curl. Total number of examples, N = 460

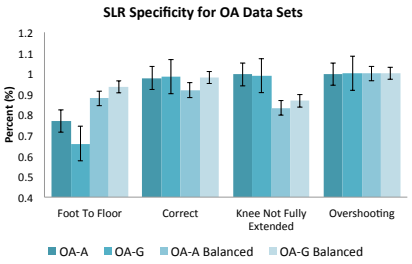
total data set. This class was detected at 100% in the OA-A data; however, that detection rate dropped to 84% when the classes were balanced. For specificity, Overshooting is 100% in all data sets (Figure 5(b)).

Table II and Table III show 10-fold cross validation results for the multi-label classifiers. There were 460 examples in the osteoarthritis data available for cross validation of the standing hamstring curl and the data included both accelerometer and gyroscope data (OA-G). The RA*k*EL method provides the best means for motion quality assessment in the SHC. The macro-specificity was 99% and the macro-sensitivity 84%. RA*k*EL can model label dependencies, which helped with the recognition.

In Table III, the RA*k*EL method again provides the best motion quality assessment for the SLR. There were 440 examples in the osteoarthritis data available for cross validation for the straight leg raise. FootToFloor was labeled in 349 of the 440 examples which contributes to the high rates of sensitivity, specificity, and even accuracy for most classifiers used. The consistently high rates (above 75%) of sensitivity for both the SHC and SLR OA-G data sets show that we are able to recognize the classes of interest in a multi-label approach.



(a) SLR Sensitivity



(b) SLR Specificity

Fig. 5: SLR results from 10-fold cross validation on OA data sets

Osteoarthritis Data with Accelerometer and Gyro (OA-G)			
	BR	ECC	RAkEL
Hamming Loss	0.03	0.04	0.03
Subset Accuracy (%)	0.88	0.87	0.90
Macro-Sensitivity (%)	0.81	0.84	0.84
Macro-Specificity (%)	0.95	0.95	0.96

TABLE III: Multi-Label Results for the Straight Leg Raise. Total number of examples, N = 440

VI. Conclusions and Future Work

We show that multi-label learning methods can be applied successfully to measure the quality of human motion data. We can pick out subtle attributes of a motion that provides information about the quality performed. As shown by the high recognition rates using the RA*k*EL method, we are also able to detect any dependencies that are present in the performance. This ability is the foundation of a system that will assess the quality of motion performed and provide feedback to the user as to the meaning of that assessment.

The results presented have some limitations, however, and there is future work to be done in order to build an in-home system. We found high rates in a laboratory environment in which we collected data for the purpose of building the learning models. For the system to work in the home, it must be capable of analyzing data in near real-time and under less controlled conditions.

Lastly, the features used in this work are among those commonly used in previous work. Although they provided a good basis to begin to assess human motion quality, additional features may be more useful.

The potential clinical impact of an intelligent rehabilitation system is significant. As the baby boomers begin, there will be a need for such systems. A low-cost system can cut healthcare costs associated with the treatment of the disease by enabling continuous and proper performance of exercises that can prolong the need for surgery or slow progression of the disease to where surgery is not needed. The techniques used for measuring the quality of human motion is a contribution to the study of how patients move with particular conditions and of assessing the quality of motion in general.

References

[1] Centers for Disease and Control Prevention. Osteoarthritis.

[2] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, 1996.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[4] E. Jovanov, A. Milenkovic, C. Otto, and P.C. de Groen. A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 2005.

[5] T. Li, C. Zhang, and S. Zhu. Empirical studies on multi-label classification. In *Proceedings of the IEEE Conference on Tools with Artificial Intelligence*, 2006.

[6] S. Melzi, L. Borsani, and M. Cesana. The virtual trainer: Supervising movements through a wearable wireless sensor network. In *Sensor, Mesh, and Ad Hoc Communications and Networks Workshops*, 2009.

[7] S.J. Preece. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. In *IEEE Transactions on Biomedical Engineering*, volume 56, 2009.

[8] P. Ravaud, P. Biraudeau, I. Logeart, J.S. Larguier, D. Rolland, R. Treves, L. Euller-Ziegler, B. Bannwarth, and M. Dougados. Management of osteoarthritis (OA) with an unsupervised home based exercise programme and/or patient administered assessment tools. a cluster randomised controlled trial with a 2x2 factorial design. *Ann Rheum Dis*, 63, 2004.

[9] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3), June 2011.

[10] R.E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 1999.

[11] A.L. Sutton, editor. *Arthritis Sourcebook*. Omnigraphics, 2010.

[12] P.E. Taylor, G.J.M. Almeida, T. Kanade, and J.K. Hodgins. Classifying human motion quality for knee osteoarthritis using accelerometers. In *IEEE Eng Med Biol Soc*, volume 2010, 2010.

[13] Y.C. Tseng, C.H. Wu, F.J. Wu, C.F. Huang, C.T. King, C.Y. Lin, J.P. Sheu, C.Y. Chen, C.Y. Lo, C.W. Yang, and C.W. Deng. A wireless human motion capturing system for home rehabilitation. In *International Conference on Mobile Data Management*, 2009.

[14] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*. Springer, 2010.

[15] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas. Mulan: A Java library for multi-label learning. *Journal of Machine Learning Research (to appear)*, 2011.

[16] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *European Conference on Machine Learning*, Warsaw, Poland, September 17-21 2007.