Journal of Industrial and Systems Engineering
Vol. 3, No. 4, pp 291-295
Winter 2010
Technical Note

JISE

Semi-Supervised Learning Based Prediction of Musculoskeletal Disorder Risk

Pankaj Chandna[1*], Surinder Deswal[2] and Mahesh Pal[3]

[1]Mechanical Engineering Department, National Institute of Technology, Kurukshetra 136119, Haryana, INDIA
pchandna08@gmail.com

[2, 3] Civil Engineering Department, National Institute of Technology, Kurukshetra 136119, Haryana, INDIA

ABSTRACT

This study explores a semi-supervised classification approach using random forest as a base classifier to classify the low-back disorders (LBDs) risk associated with the industrial jobs. Semi-supervised classification approach uses unlabeled data together with the small number of labelled data to create a better classifier. The results obtained by the proposed approach are compared with those obtained by a backpropagation neural network. Comparison indicates an improved performance by the semi-supervised approach over the random forest classifier as well as neural network approach. Highest classification accuracy of 78.20% was achieved by the used semi-supervised approach with random forest as base classifier in comparison to an accuracy of 72.4% and 74.7% obtained by random forest and back propagation neural network approaches respectively. Thus results suggest that the proposed approach can successfully classify jobs into the low and high risk categories of low-back disorders based on lifting task characteristics.

***Keywords*:** low-back disorders, semi-supervised learning, backpropagation neural network, random forest classifier

1. INTRODUCTION

The term musculoskeletal disorders (MSDs) refer to conditions that involve the nerves, tendons, muscles, and supporting structures of the body. Work related Musculoskeletal disorders (WMSDs) refer to musculoskeletal disorders to which the work environment and the performance of work contribute significantly or that are made worse or longer lasting by work conditions. WMSDs are among the most prevalent lost-time injuries and illnesses in almost every industry (Bureau of Labor Statistics (1996); Tanaka et al. (1995)) and specifically those involving the back, are among the most costly occupational problems (Guo et al. (1995); Frymoyer and Cats-Baril (1991); Webster and Snook (1994)). They may cause a great deal of pain and suffering among afflicted workers and may decrease productivity and the quality of products and services. Literature review suggests a strong relationship between selected MSDs of the upper extremity and the low back and exposure to physical factors at work. Specific attention is given to analyze the weight of the evidence for the strength of the association between these disorders and work factors. Because the relationship between exposure to physical work factors and the development and prognosis of a particular

* Corresponding Author

disorder may be modified by psychosocial factors, the literature about psychosocial factors and the presence of musculoskeletal symptoms or disorders is also reviewed. Understanding these associations and relating them to the cause of disease is critical for identifying exposures amenable to preventive and therapeutic interventions. A substantial body of credible epidemiologic research provides strong evidence of an association between MSDs and certain work-related physical factors when there are high levels of exposure and especially in combination with exposure to more than one physical factor (e.g., repetitive lifting of heavy objects in extreme or awkward postures).

The modelling techniques may be used for the development of models that explicitly describe the risk associated with various work designs so that specific and quantitative workplace assessments can be made (Dempsey and Westfall (1997)). Although there are established risk factors for LBDs, the manner in which these factors interact to promote the risk of LBDs in industry, and ultimately disability, is not well understood (Zurada et al. (1997)). Many researchers have demonstrated the use of advanced statistical methods including logistic regression and generalized additive models (Dempsey et al. (1995); Marras et al. (1993)) as well as artificial neural networks (ANNs) (Karwowski et al. (1994); Killough et al. (1995); Zurada et al. (1997)) to predict musculoskeletal disorder risk associated with occupational exposures. It is found that neural networks provide superior predictive capability in comparison to multiple linear regression models (Killough et al. (1995); Zurada et al. (1997)).

Recently, studies suggest that it is beneficial to use semi-supervised classification approach in situations where number of labeled data are sparse. The goal of semi-supervised classification is to use un-labeled data to improve the generalization (Seeger (2002)). Thus, the main objective of this study is to use a semi-supervised classification approach (Driessens et al. (2006)) that uses random forest as base classifier to classify industrial jobs according to the potential risk for low back disorders (LBDs). Such a system could be quite useful in hazard analysis and injury prevention due to manual handling of loads in an industrial environment.

2. SEMI-SUPERVISED LEARNING

Semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data during training. Semi-supervised learning is a combination of both unsupervised and supervised learning approaches. Researchers in the field of machine learning suggest that use of a large number of unlabeled data, when used in combination with a small number of labeled data, can improve the classification accuracy (Zhu (2006)). A number of approaches for semi-supervised learning are proposed in literature. These include the use of expectation-maximization algorithm, self-training, co-training, transductive support vector machines, split learning, and graph-based methods. For further details about these approaches readers are referred to Zhu (2006).

This study uses a semi-supervised approach recently proposed by Driessens et al. (2006). This classifier uses both labeled and unlabeled data in a two-stage set-up. In the first stage a standard classifier (random forest in the present study) is used with the available training data. In the second stage, the model generated from the training data is used to classify all the examples in the test set. These classified test data are then used together with the original training data in a weighted classification algorithm. The weights used by the classifier are meant to limit the amount of trust the algorithm puts in the labels generated by the model from the first step. As a default value, the weights of the training data are set to 1 and the weights of the pre-labeled test-data to N/M with N and M are the number of training and test data respectively. A parameter F is also added to the algorithm so as to set the weights of the test data sets equal to $F \times (N/M)$. This step makes it possible to vary the influence given to the unlabeled data as well as the classifier built in step1. Values of F

between 0 and 1 will lower the influence on the test-data and the learned model from the first step whereas values larger than 1 will increase their influence.

3. RANDOM FOREST CLASSIFIER

Random forest classifier consists of a combination of tree classifiers where each classifier is generated using a random vector sampled independently from the input vector, and each tree casting a unit vote for the most popular class to classify an input vector (Breiman, 1999). Random forest classifier used for this study consists of using randomly selected features or a combination of features at each node to grow a tree. Bagging, a method to generate a training data set by randomly drawing with replacement *N* examples, where *N* is the size of the original training set (Breiman, 1996), was used for each feature/feature combination selected in the design of random forest. Each bootstrap training set consists about 67% of data from original training set thus about one-third of the data are left out from every tree grown. This left out data set is called out-of-bag (out of the bootstrap sample) and can be used as a test set for the tree grown on the bootstrap sample. Design of a decision tree required the choice of an attribute selection measure and a pruning method. Random forest classifier uses the Gini Index as an attribute selection measure, which measures the impurity of an attribute with respect to the classes.

In a random forest classifier a tree is grown to the maximum depth on new training data using a combination of features. These full-grown trees are not pruned. This is one of the major advantages of random forest classifier over other decision tree methods like the one proposed by Quinlan (1993). As the studies suggest that the choice of the pruning methods and not the attribute selection measures, affect the performance of tree based classifiers (Pal and Mather (2003)). Breiman (1999) suggests that as the number of tree increases, the generalisation error always converges even without pruning the tree and overfitting is not a problem because of the Strong Law of Large Numbers (Feller (1968)). Number of features used at each node to generate a tree and the number of trees to be grown are two user-defined parameters required for a random forest classifier. At each node, only selected features are searched through for the best split. Thus, the random forest classifier consists of *N* trees, where *N* is the number of trees to be grown which can be any user defined value. To classify a new data set, each case of the data set is passed down to each of the *N* trees. The forest chooses a class having the most out of *N* votes, for that case.

4. DATASET AND METHODOLOGY

The data set collected by Marras et al. (1993) in a field study of 235 industrial jobs with low- and high-risk values of low-back disorders was used in this study. This consists of independent variables namely lift rate in number of lifts per hour (LIFTR), peak twist velocity average (PTVAVG), peak moment (PMOMENT), peak sagittal angle (PSUB), and peak lateral velocity maximum (PLVMAX) and a single dependent variable (i.e. low-back disorders) having two discrete levels of ‘low risk' and ‘high risk’. A detailed table of both training and test data is provided in Zurada et al. (1997). They used a backpropagation neural network to classify this data set according to the potential for low-back disorders. They used a total of 148 randomly selected data (i.e. 74 numbers of low-risk and 74 numbers of high-risk jobs) for training the neural network. The remaining 87 jobs (50 low risk and 37 high-risk) were used to test the neural network. In order to compare the performance of the proposed approach with Zurada et al. (1997), same data set for training and testing was used in this study. Here class 1 refers low-risk jobs whereas class 2 represents high-risk jobs. The proposed semi-supervised approach for LBDs classification using random forest as base classifier was compared with random forest as well as a backpropagation neural network classifier in terms of overall as well as individual class accuracies. With random

forest classifier, two numbers of features at each node to generate a tree and the 50 number of trees was found to be working well with this data set.

5. RESULTS

The proposed semi-supervised approach with random forest as base classifier was able to classify a total number of 68 jobs out of 87 correctly (78.2%). Out of 50 low-risk and 37 high-risk jobs, a total of 40 low-risk and 28 high-risk jobs were correctly classified respectively (Table 1). Comparison of results provided in table 1 suggests that the proposed semi-supervised classification approach increases classification accuracy by more than 7% in comparison to a simple random forest classifier.

Table 1 Classification accuracies with random forest and semi-supervised random forest based classification approach

	Random forest		**Semi-supervised random forest**	
Classifier	**Class 1**	**Class 2**	**Class 1**	**Class 2**
Class 1	34	16	40	10
Class 2	9	28	9	28
Class accuracies	68%	75.7%	80%	75.7%
Overall classification accuracy	71.3%		78.2%	

An increase of 3.5% in overall classification accuracy was achieved by the semi-supervised classification approach in comparison to the neural network classifier, suggesting a better performance by the proposed approach with this data set. Further, proposed approach provides an improved accuracy in classifying low risk jobs as compared to the neural network approach. A slight decrease in accuracy is observed with the high risk jobs, which may be due to the reverse placement of the data as mentioned in Zurada et al. (1997).

6. CONCLUSION

This study discusses a semi-supervised classification approach with random forest as base classifier to classify low-back disorders (LBDs) risk associated with industrial jobs. Results suggests that the proposed approach works quite well in classifying the low-back disorder (LBDs) risk associated with industrial jobs in comparison to the neural network approach as proposed by Zurada et al. (1997) in terms of overall as well individual class classification accuracy.

REFERENCES

[1] Breiman L. (1996), Bagging predictors; *Machine Learning* 26; 123-140.

[2] Breiman L. (1999), Random forests-Random Features; Technical Report 567; Statistics Department, University of California, Berkeley, http://ftp.stat.berkeley.edu/pub/users/breimanf,4 July 2007.

[3] Bureau of labour statistics (1996), Characteristics of injuries and illnesses resulting in absences from work, 1994 Washington, DC: US Department of Labour; *Bureau of Labour Statistics, USDL*; 96–163.

[4] Dempsey P.G., Ayoub M.M., Westfall P.H. (1995), The NIOSH lifting equations: a closer look. In: Bitner A.C. and Champney P.C. (Ed); *Advance in Industrial Ergonomics and Safety VII. Taylor and Francis, Bristol, PA*; 705–712.

[5] Dempsey P.G., Westfall P.H. (1997), Developing explicit risk models for predicting low-back disability: a statistical perspective; *International Journal of Industrial Ergonomics* 19; 483–497.

[6] Driessens K., Reutemann P., Pfahringer B., Leschi C.(2006), Using Weighted Nearest Neighbour to Benefit from Unlabeled Data.In:Wee Keong Ng,Masaru Kitsuregawa, Jianzhong Li, and Kuiyu Chang, (Ed); *Advances in Knowledge Discovery and Data Mining, 10th Pacific-AsiaConference, PAKDD 2006, 3918 of LNCS*; 60-69.

[7] Feller W. (1968), An Introduction to Probability Theory and Its Application; third edition, Vol. 1, Wiley; New York.

[8] Frymoyer J.W., Cats-Baril W.L. (1991), An overview of the incidence and costs of low-back pain; *Orthopaedic Clinics of North America* 22; 262–271.

[9] Guo H., Tanaka S., Cameron L., Seligman P., Behrens V., Ger J. (1995), Back pain among workers in the United States: national estimates and workers at high risks; *American Journal of Industrial Medicine* 28; 591–602.

[10] Karwowski W., Zurada J., Marras W.S., Gaddie P. (1994), A prototype of the artificial neural network-based system for classification of industrial jobs with respect to risk of low back disorders. In: Aghazadeh F. (Ed); *Advances in Industrial Ergonomics and Safety VI, Taylor & Francis, Bristol, PA*; 19–22.

[11] Killough M.K., Crumpton L.L., Calvert A., Bowden R. (1995), An investigation using neural networks to identify the presence of carpal tunnel syndrome; *4th Industrial Engineering Research Conference Proceedings, IIE, Norcross, GA*; 659–667.

[12] Marras W.S., Lavender S.A., Leurgans S., Sudhakar L.R., Allread W.G., Fathallah F., Ferguson S. (1993), The role of dynamic three dimensional trunk motion in occupationally-related low back disorders; *Spine* 18; 617–628.

[13] Pal M., Mather P.M. (2003), An assessment of the effectiveness of decision tree methods for land cover classification; *Remote Sensing of Environment* 86; 554–565.

[14] Quinlan J.R. (1993), C4.5: Programs for Machine Learning; Morgan Kaufmann, San Mateo.

[15] Seeger M. (2002), Learning with labeled and unlabeled data; Technical report, Institute for Adaptive and Neural Computation, University of Edinburgh, http://www.kyb.tuebingen.mpg.de/bs/people/seeger/papers/review.pdf, 13 May 2007.

[16] Tanaka S., Wild D., Seligman P., Halperin W., Behrens V., Putz-Anderson V. (1995), Prevalence and work-relatedness of self-reported carpal tunnel syndrome among US workers: Analyses of the occupational health supplement data to the 1988 National Health Interview Survey; *American Journal of Industrial Medicine* 27; 451–470.

[17] Webster B.S., Snook S.H. (1994), The cost of compensable upper extremity cumulative trauma disorders; *Journal of Occupational and Environmental Medicine* 36; 713–717.

[18] Zhu X. (2006), Semi-supervised learning literature survey; Technical Report 1530, Department of Computer Sciences, University of Wisconsin-Madison, http://www.pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf, 8 May 2007.

[19] Zurada J., Karwowski W., Marras W.S. (1997), A neural network-based system for classification of industrial jobs with respect to risk of low back disorders due to work-place design; *Applied Ergonomics* 28; 49–58.