IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555
Vol.6, No.2, Mar-April 2016

Novel approach for nutrition activities classification By clustering with semi-supervised learning

Harpreet Singh
Student, Lovely Professional University
Phagwara, Punjab

Sonal Arora
Asst. Professor, Lovely Professional University
Phagwara, Punjab

ABSTRACT

Semi-supervised learning (SSL) is an important research problem in machine learning. While it is usually expected that the use of unlabeled data can improve performance, in many cases SSL is outperformed by supervised learning using only labeled data. To this end, the construction of a performance-safe SSL method has become a key issue of SSL study. In this paper classified the effect of fast-food on human body by clustering with supervised learning and improve the clustering. This paper also use feature selection and feature extraction.

INTRODUCTION

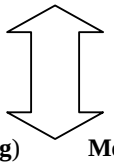
Eat appropriately and live healthy is the mandatory prerequisites for long life. Tragically, today most of the public has been adjusted to a consumption of food which has negative impact on the human body. Life is very fast nowadays so one has very less time to truly think what he/she is eating is correct! Globalization and urbanization are the big reasons which have affected ones eating habits and compel the people to take food which contain excessive calories, famously known as '"junk foods ". Cases related to the diseases i.e. coronary vein illness and diabetes mellitus are rising significantly in number in developing nations and the only reason for this is consumption of fast food. This research paper develops a model which includes semi-supervised learning to collect data. Semi-supervised learning consists of a small portion of labelled data but on the other hand it contains of a big portion of unlabeled data. Semi-supervised learning stands in between unsupervised learning (training data that is not labelled) and supervised learning (training data that is labelled). Researchers have discovered when unlabeled data is used with a small portion of labelled data; it can generate a significant enhancement in learning accuracy. The collection of labelled data for a machine learning problem generally needs an experienced expert. The cost factor related to labelled training set makes it infeasible most of the times, whereas collection of unlabeled data is comparatively not an expensive task. In such scenarios, semi-supervised learning has its own considerable importance i.e. it has more practical value. The information derived in Semi-supervised learning through the combination of both is passed to the classification part. How good the classification is that could be achieved either by eliminating the unlabeled data (supervised learning) or by eliminating the labels (unsupervised learning).Semi-supervised learning is also denoted either by transductive learning or inductive based learning. Transductive learning refers to deduce the right labels for a given unlabeled data $x_{l+1}, \cdots, x_{l+u}$ only. Inductive learning refers to deduce the right mapping i.e. from X to Y. In the next part, feature extraction, some features are extracted, and after that observing those features which are very much suitable to attain an efficient classification. Clustering is generally an unsupervised criterion. With the help of clusters we can collect same items in one cluster i.e. items which have same properties. So in this way we can make several clusters in which each cluster contain items with similar properties. We can classify each cluster with the help of classification. This classification will be based on clustering information derived from class. The scheme which has derived from this is used to forecast heart disease also it produces the efficient classification mechanism related to multidimensional data .Clustering helps to reduce the dimensions to reduce the error in classification.

REVIEW OF LITERATURE SURVEY

Francisco Duarte , Andre Lourenco, Arnaldo Abrantes(2014),"Classification of physical activities using a smart phone: evaluation study using multiple users" defines that physical inactivity is a big problem with which population of developed countries is affected. Depression, obesity etc are the consequences of it. This work was helpful to remove these problems .This study was aimed to motivate a person and to monitor physical activity .This was only possible through accelerometer sensor. To accurately

classify activities this research work had carried out very deep analysis of the acquired sensors signal. From these acquired signal it was easy to know the most distinctive features. To get the signals properly , the phone was placed along the waist in the front pocket of right side. This work had helped to get the proper idea about the superior method for the classification of activities. Accelerometer used here as a solution for monitoring the activities performed by user. This approach was helpful in promoting the more productive lifestyle.

Efficient activity classification



(by knowing) Most relevant features + Algorithm

Figure1: Brief idea for Activity Classification

In the beginning part, the signal is procured amid the movement utilizing the cell phones accelerometer sensor. In the second module , feature extraction, a few features are separated, both in time and frequency space, and afterward investigated to permit those features that are best suited to achieve an accurate classification. features were extracted for both time and frequency domain, to know which is important one or not.

M.I. López, J.M Luna, C. Romero, S. Ventura ," Classification via clustering for predicting final marks based on student participation in forums " includes the prediction of marks for students in a university course. They have used classification via clustering approach to get the good results. Here they include the concept of forums. Forums are the tools in online learning environments Here students discussions are there related to university course it means it facilitates collaborative learning. The goal of this paper is twofold: to find out if some student makes regular discussions on forum data can it be used for the prediction of the final scores for the university subject? The other goal is to check whether the newly find out classification via clustering approach can provide similar accurate results as that of old clustering algos. The data is gathered from the first year students. It is a real data. Numerous clustering algorithms was implemented with the proposed approach after that they were compared with old classification algorithms to predict whether the pupils fail or pass the subject. The basis of this prediction is forum data. For the group of selected attributes Expectation-Maximization clustering algo produces similar results as that of best classification algorithms, only for the set of selected attributes. At last relationship between two clusters is find out with the help of centroids. Centroids play a key role in the aspect of clustering. If the centroid and the point of one cluster lies farther from each other then more will be the dissimilarity. But if distance is less then less will be the conflicting properties.

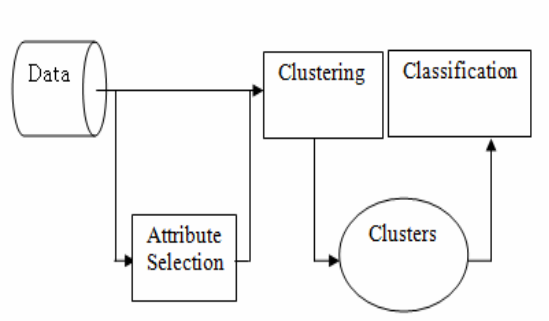


Figure 2: Classification via Clustering

Mohd Fikri Azli bin Abdullah, Ali Fahmi Perwira Negara, Deok-Jai Choi, N Kalaiarasi Sonai Muthu,"Classification Algorithms in Human Activity Recognition using Smartphones" discusses the various classification algorithms. The algos which are used in smartphone based human activity recognition. Unprocessed data from sensors and devices includes a lot of meaningful data and also large amount of unwanted data . Now the challenge arises from this is related to the extraction of features. It is the only way to get rid from the noisy data. With the help of the feature extraction we can select the convenient meaningful data from the availabe unrefined data. If we have only the meaningful data then classification criteria will take less time and memory also the performance will boost up. Perfection in classification is achieved through increasing the number of features. With this weight age of one class will increase as more values are there to support it. But it will lead to the more consumption of time and memory. Features are grouped into four parts. About 80% of works depends upon the magnitude value so magnitudinal features are in great demand because it reduces the complexity. Classification algo plays an important role in human activity recognition. How to select the classification algo? It depends upon the potential of the processing node to implement the algorithm. Besides this, the performance of the computation algorithm is estimated through some computation process. Large number of researchers use supervised classification algorithms. To construct classification model, firstly the algo is trained with data which contain labels. After that this mechanism is used to classify the newly incoming data. Survey tells, support vector machines, K nearest

neighbor, Neural network, decision trees etc are the algos which are mostly used. Brent Longstaff et al. conducted their in-depth research on two semi-supervised learning processes First one is Self-learning consist of one classifier on the whole. Prediction on the behalf of self - learning's classifier if comes under the range of higher level, labelize the data with that particular prediction. Second one is co-learning which includes more than one classifiers. There is a great need that implementation should be done in servers in the case of supervised classification algorithms. Because it evolves complex computation in which model is derived from training data. KNN is a very good algorithm known as instance based. It is used to classify human actions according to the individual input. So these types of algos proves to be a boon to be executed in smartphone, one of the reasons for this is that it requires less time and memory resources.

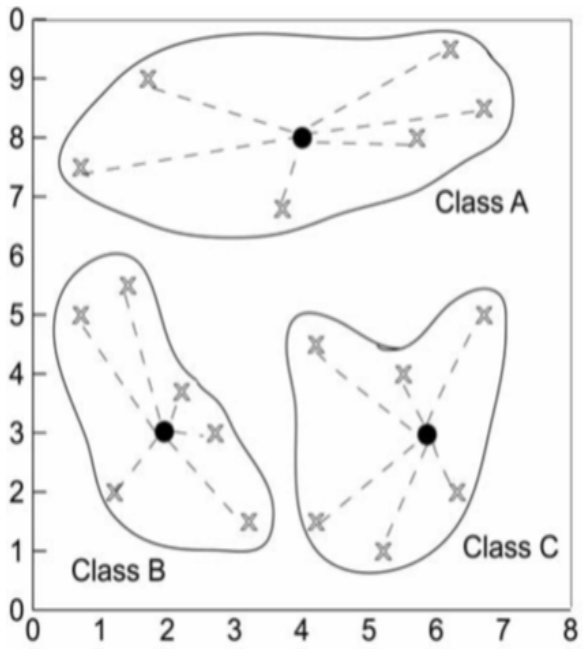


Figure 3: Classification done with the help of clustering

Xiaojin Zhu,"Semi-Supervised Learning Literature Survey" discusses various aspects of semi-supervised learning. Labeled data are very much difficult, expensive, or time consuming to obtain. . Semi-supervised learning deals with this issue by using large percentage of unlabeled data, together with the little amount of labeled data, to build good classifiers. Semi-supervised learning is advantageous covers less human effort and provides higher accuracy. What number of semi-supervised learning methods are there? Many. Some frequently-used methods include: EM with generative mixture models, co-training, transductive support vector machines, self-training and graph-based methods. One should use a criteria whose suppositions fit the problem structure. This may be tough in real circumstances. To deal with toughness we can try the following : Do the classes generate very good clustered data? In the case of yes, EM with generative mixture models is a good option; Do the features naturally break into two groups? In the case of yes, co-training is appropriate choice; Do the two points with similar features tend to be assumed in the same class? In the case of yes, graph-based methods may be appropriate; using SVM? Transductive SVM is a good choice.

N. Elavarasan , Dr. K.Mani," A Survey on Feature Extraction Techniques" defines that there is a large amount of data available. Therefore we need some technique to extract meaningful information i.e. knowledge from it. The name of that technique is data mining . But the data obtained from data mining have high dimensionality. It is a big problem. Machine learning algorithms cannot survive better because of this problem efficiency decreases. The important way to get rid of this problem is dimensionality reduction. DR serves as an important part of data mining. So in dimensionality reduction most effective technique is feature extraction. In this way effective classification can be achieved . Dimensionality reduction is an useful preprocessing technique. We have large amount of data from which if we want can extract features and assure good classification. But it will create problem in producing the classifiers also the dimensions are at higher level and surely it will effect the classification badly and produces error. so reduce the dimensions before pursuing for classification. Dimensionality reduction helps data mining with this issue.

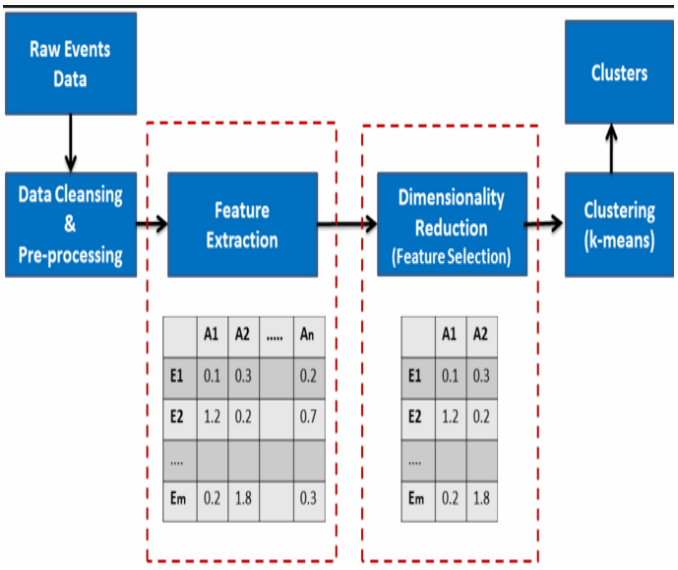


Figure 4: Dimensionality reduction

BACKGROUND

In the previous work i.e." Classification of physical activities using a smart-phone " only supervised learning has been used . They used their dataset and classifiers to obtain the results. Four different classifiers have used in that paper to strengthen the results. Now we are going to use the semi-supervised learning to develop a new efficient model on different dataset because it is very difficult to arrange same dataset on which previous work has been done . Also classification with the help of clustering will strengthen our proposed model. In this paper," Fast Foods and their Impact on Health " researcher discussed the various impacts of junk food on our health so with the help of it this paper provides a message to people to leave eating junk food because excessive calories will generate various negative effects in our body.

PROPOSED METHODOLOGY

Primarily the data needs to be gathered. Because that data is a rough data so there is a great need to preprocess it. It will reduce the size of data so it saves time during processing. Because we are going to use semi-supervised classification approach there is a great requirement of classifier i.e. meta classifier , on the basis of which classification will be conducted. This process is based on the supposition i.e. every cluster denotes some class. To decrease the dimensionality of data we have to concentrate on important attributes only not all the attributes. So optional attribute selection process is profitable here. Now we have the little refined training data from which we will extract the features. In the next part, we will implement the clustering algorithm using the featured data. Clustering will make several clusters and each cluster will contain several items. Now it is a turn of mapping between clusters and classes. From that mapping it will be easy to predict the class labels i.e. features with labels. During clustering we cannot use class attribute but after the clustering it is worthful because it will become the classifier of a particular cluster. It is good if the clusters generated is same in number as of class labels because it helps to derive an excellent model. After the new model is developed we will test it that it is providing better results than old model or not. Better clustering will lead to better classification otherwise the scenarios will arise which produce erroneous results. Extraction of features will play key role here.

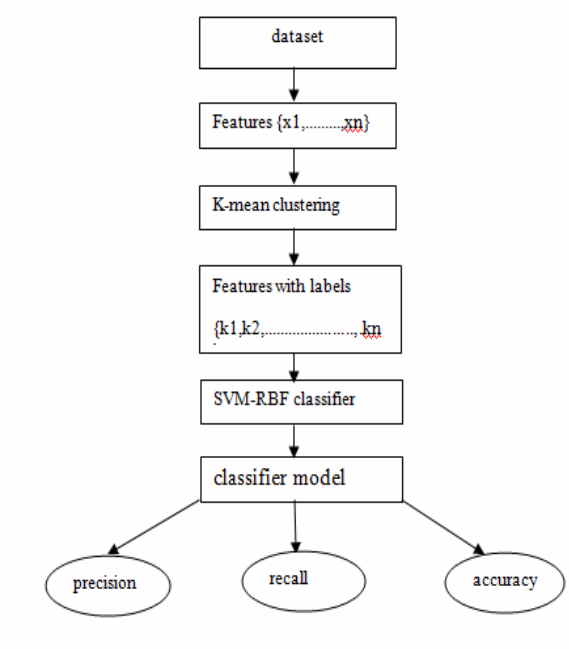


Figure 5: flow chart representing brief methodology

CONCLUSION

Semi-supervised learning plays a vital role in machine learning as it is a cost effective solution against labeled data because it includes large amount of unlabeled data than labeled one. The purpose of this article is to develop an efficient model to draw good results based on different classifiers. Efficient classification is only achieved through good clustering algorithms. Also this research work provides awareness to the user against junk food. In the near future this research work will become basis of several medical field applications. It will be possible to draw the attention of researchers towards the medical field to serve the society.

REFERENCES

[1] Francisco Duarte , Andre Lourenco, Arnaldo Abrantes ," Classification of physical activities using a smartphone ; evaluation study using multiple users (2014)" Procedia Technology 17 (2014) 239 – 247 Conference on Electronics, Telecommunications and Computers – CETC 2013.

[2] (IJARAI) International Journal of Advanced Research in Artificial Intelligence,
Vol. 2, No.2, 2013

[3] Ashakiran & Deepthi ," Fast Foods and their Impact on Health ." JKIMSU, Vol. 1, No. 2, July-Dec. 2012. Department of Biochemistry, 2Department of

Community Medicine, Sri Devaraj Urs Medical College,Kolar-563101 (Karnataka), India.

[4] Classification via clustering for predicting final marks based on student participation in forums M.I. López, J.M Luna, C. Romero, S. Ventura Department of Computer Science and Numerical Analysis University of Córdoba
Córdoba, Spain.(2012)

[5] Mohd Fikri Azli bin Abdullah, Ali Fahmi Perwira Negara, Md. Shohel Sayeed, Deok-Jai Choi, Kalaiarasi Sonai Muthu (2012) ," Classification Algorithms in Human Activity Recognition using Smartphones". World Academy of Science, Engineering and TechnologyVol:6 2012-08-27.

[6] S. Jyoti, A. Ujma, S. Dipesh, and S. Sunita. Predictive data mining for medical diagnosis: An overview of heart disease
prediction. *International Journal of Computer Applications*, 17(8):43–48, 2011.

[7] R. Krakovsky and R. Forgac. Neural network approach to multidimensional data classification via clustering. Intelligent Systems and Informatics (SISY), 2011 IEEE 9th International Symposium on, 169–174, IEEE2011.

[8] M. Panda and M. Patra. A novel classification via clustering method for anomaly based network intrusion detection system. *International Journal* of Recent Trends in Engineering, 2:1–6, 2009.
[12] R. Rabbany, M. Takaffoli

[9] N. Elavarasan , Dr. K.Mani," A Survey on Feature Extraction Techniques" International Journal of Innovative Research in Computer and Communication Engineering

[10] C. Romero, S. Ventura, P. Espejo, and C. Hervás. Data mining algorithms to classify students. Proceedings of Educational Data Mining, 20-21, 2008

[11] Semi-Supervised Learning Literature Survey Xiaojin Zhu (2007).