International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 09 | Sep-2016 www.irjet.net p-ISSN: 2395-0072

An Accessible Prediction System for Complex Statistics: A Survey

Akash Sharma[1],Nikita Jain[2]

[1]*M.Tech Scholar, Department of Computer Science & Engineering, GIT, JAipur,Rajasthan,India*
[2]*Assistant Professor, Department of Computer Science & Engineering, GIT, JAipur,Rajasthan,India*

Abstract – ***Over the past few decades due to some significant characteristics the field of design an expert prediction system has attains huge attention of the researchers. An efficient prediction system may helpful to predict an event more efficiently and effectively which reduce the cost, time and effort. Depending on different data and analysis technique a number of prediction systems have been proposed by numerous researchers but due to non availability of any single effective method to handle several territories the field still have a scope of modernization. This paper presents a state of art work on accessible prediction systems which may helpful for the researchers to design and implement an effective scheme by considering the current issues which are associated with an on-hand prediction techniques.***

Key Words: *Data mining techniques, decision tree, Classifier, NaiveBayes.*

1. INTRODUCTION

Nowadays, with the technological advancement the uses of digital information based system have got a massive popularity in near about each and every area of work. During recent ages, a growing amount of peoples use these systems as a source of information and there is like to be impractical for an individual and organizations to accomplish their daily tasks without relying on the conveniences provided by these systems. Although still users have doubts about truthfulness of the proliferate contents. In some crucial working fields such as in the area of medical where disease may the cause of death the incorrect or deferred predicted information may cause of danger or more harmful to a person, therefore quickly and accurately prediction of syndrome signs in time is must for shaping the trust. On the other hand, due to various causes the software appears with many defects. Predicting defects of an application at an early stage minimize the cost, time and pick up the overall effectiveness, attain a huge growth in recent days [1, 2]. Additionally precious information is always prone over the network. Therefore quickly and accurately finding of the content that influences the trust is very important.

This paper presents the current research status and the key techniques of existing prediction approaches with the associated issues. The rest of paper is organized as follows. Prediction system and the techniques use in the on-hand approaches are elaborate in section 2. Section 3 presents the related work and the issues associated with on-hand prediction system is present in section 4. Finally section 5 concludes the paper.

2. PREDICTION SYSTEM & TECHNIQUES

Technically, the term predictions are based on the accuracy of verdict connection or patterns among dozens of fields in large databases. For such type of achievement the regular process of database updating and machine training are essential key features. The figure 1 depict a learning process

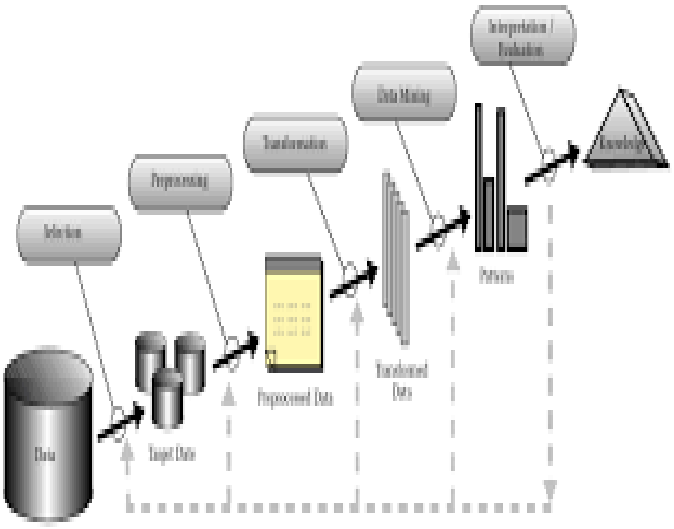

Fig. 1 Knowledge Discovery in Databases

Usually, for a machine learning process, two foremost components, supervised and unsupervised learning are most popular routines [5-8].

2.1 Supervised Learning Technique

Usually Supervised learning technique is a machine learning (Data Mining) task from supervised training data.

The technique execute in a regulation to generalize from the training data to correctly determine the class labels for unseen instances. It offer training example where each example is a pair, consisting of an input object and a desired output value also called the supervisory signal. The algorithm examines grounding statistics and constructs an inferred function, which is called a classifier if the output is discrete or a regression function if the output is continuous [9, 10]. This technique executes following steps to resolve a certain dilemma.

1. Before start execution processes builds up a set of the tag data and training examples for system training.
2. Organized a formation of training function and equivalent learning algorithms with the key feature.
3. Carry out the training process with the collected training sets and supervised methods.
4. Assess outcomes in term of accuracy of the system.
5. Evaluate system routine with a different statistics from the training sets.

The figure 2 depicts the process of supervised learning model.

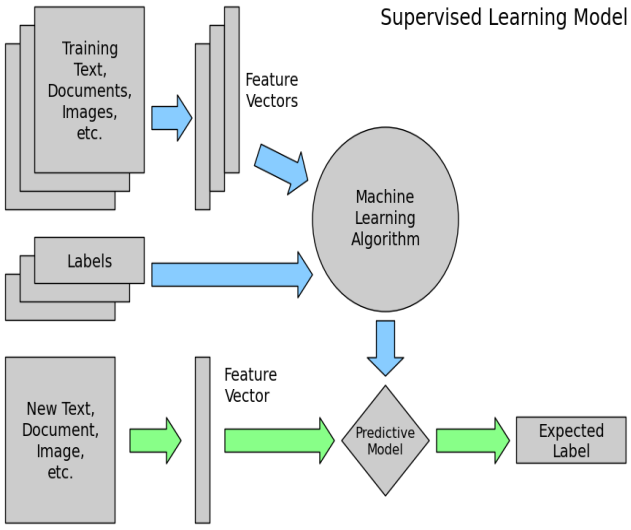

Fig. 2 Supervised Techniques

2.2 Unsupervised Learning Technique

Unlike supervised technique the unsupervised methods trying to find hidden structure in unlabeled data. Unsupervised learning is important since it is likely to be much more common in the brain than supervised learning. The system learns in a state where any previous release is not built and represents particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns.

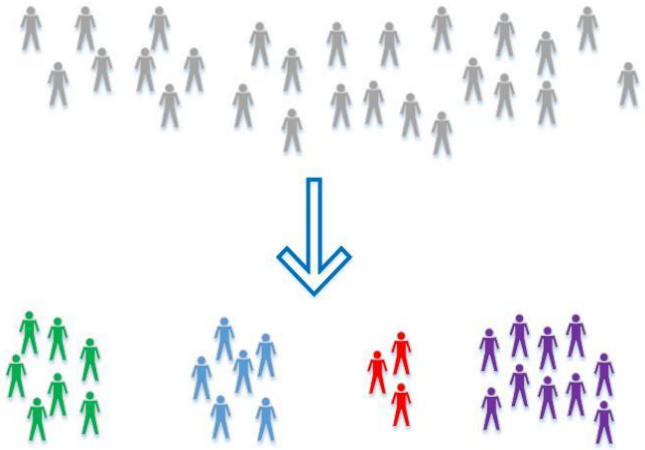

Fig. 3 Unsupervised Techniques

3. RELATED WORK

A large amount of preceding work has been done in this field, ranging from novel algorithms to literature reviews [11–15]. To analysis the diabetes disease a comparative investigation has been present in [16]. The approach use two most classification algorithms, naïve bayes and decision tree with same dataset for finding the effective predictive method. A result of decision tree presents its efficiency over the other uses classification technique. In same context another approach [17] uses J48, Random Forest, Naive Bayes algorithms for calculating the algorithms prediction accuracy level. The results were same as the previous approach.

The paper [18] presents a comparative investigation on the base of multi classification algorithms. The author use Naive Bayes, decision tree (J48), Sequential Minimal Optimization (SMO), Instance Based for K-Nearest neighbour (IBK) and Multi-Layer Perception with the 3 different dataset of breast cancer to find the classification accuracy. The approach has analysis the uses algorithm on the base of 10 folds cross validation technique. A combination at classification level is accomplished between these classifiers to get the best multi-classifier approach and accuracy for each data set.

Diabetes and cardiac diseases [19] are predicted using Decision Tree and Incremental Learning at the early stage. The i+Learning and i+LRA performs better than ID3 and other incremental learning algorithms on the bases of classification accuracy. These both algorithms can even handle the new attributes without affecting the learning performance. The main drawback of this method is that it adopts the binary tree rather than multi-branch tree. Numerous authors have used different techniques for design an effective prediction system. To classify radar images level base algorithm proposed in [20]. The proposed approach has divides the images into segments

at its first level and use the segmented part for classification process at its second level. In same context another approach [21] has present to classifying meteorological volumetric radar data in order to detect storm events responsible for summer severe weather.

The author in [22] has use support vector clustering (SVC)-based probabilistic approach for unsupervised chemical process monitoring and fault classification. A number of dataset has been used with the binary encoded output based data weighting (BEOBDW) technique for generalize the proposed data weighting method. To classify the imbalanced data an approach presents in [23, 24] use SVM classification technique. The authors of [25] proposed a twin support vector machine (TWSVM) in order to solve the shortages of the large calculation amount and slow classification speed of SVM. And an automatic classification method of star spectra data based on manifold fuzzy twin support vector machine (MF-TSVM) is proposed. The approach present in [26] proposed a novel fast feature selection method based on multiple SVDD and applied it to multi-class microarray data.

4. CURRENT ISSUES WITH ON-HAND PREDICTION SYSTEM

Since the age of design a prediction system numerous approaches have been proposed by using different techniques. However the approaches enhance the accuracy of the traditional prediction system but still field face an issues of low classification accuracy and speed. Additionally major accessible approaches use single classification algorithm for the purpose of classifying data. Therefore the approaches have their own limitations due to no one single classification algorithm is enough to classify different data sets. The major issues which are associated with the current prediction system can be point out as.

1. Major systems are designed with supervised learning techniques. Therefore they are inefficient to predict anomaly event accurately.
2. The accuracy and speed of present system is not effective, require modernization.
3. A good number of systems are designed with the strength of single classification technique. Therefore they have own limitation with the classifying datasets.
4. Inefficient to classify with real traffic. Work efficiently and effectively with the data and feature set at which they are tuned and trained.
5. Major of approaches produces huge false alarms.

5. Conclusion & Future Work

Since the age of an information system a number of prediction techniques have been proposes by different researchers with the aim to enhance the accuracy of on-hand systems. This paper presents an up to date investigation on prediction system and its accessible techniques. After presenting a good amount of related work this paper has also present several of issues which are still associated with the on-hand prediction mechanism. Different approaches work in a different way according to the assumptions and employs mechanisms. Therefore the approaches have different strengths and drawbacks. These inadequacies of on hand prediction mechanism motivate us to make an endeavor in future to design a new effective and efficient system which can predict a statistics more accurately and speedily.

REFERENCES

[1] Gerry Coleman and Renaat Verbruggen. A quality software process for rapid application development. Software Quality Control, 7(2):107–122, 1998.
[2] Daniel Dawson, Nathan Hawes, Christian Hoermann, Nathan Keynes, and Cristina Cifuentes. “Finding bugs in open source kernels using parfait” Sun Microsystems, November 2009.
[3] K. Thenmozhi, P.Deepika “Heart Disease Prediction Using Classification with Different Decision Tree Techniques” International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014
[4] Qasem A. Al-Radaideh, Eman Al Nagi “Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012.
[5] Kotsiantis, S. B. "Supervised Machine Learning: A Review of Classification Techniques." Informatica, 2007: 249-268.
[6] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch. Benchmarking classification models for software defect prediction: A proposed framework and novel findings. IEEE Trans. Softw. Eng., 34(4):485–496, 2008.
[7] T. Menzies, J. Greenwald, and A. Frank. Data mining static code attributes to learn defect predictors. IEEE Transactions on Software Engineering, 33(1):2–13, January 2007.
[8] T. Menzies, B. Turhan, A. Bener, G. Gay, B. Cukic, and Y. Jiang. Implications of ceiling effects in defect predictors. In Proceedings of the 4th international workshop on Predictor models in software engineering, New York, NY, USA, 2008. ACM.
[9] N. Nagappan and T. Ball. Static analysis tools as early indicators of pre-release defect density. In ICSE 2005, St. Louis, 2005.

[10] T. J. Ostrand, E. J. Weyuker, and R. M. Bell. Predicting the location and number of faults in large software systems. IEEE Transactions on Software Engineering, 31(4):340–355, 2005.
[11] Norman E. Fenton, Martin Neil, Ieee Computer Society, and Ieee Computer Society. A critique of software defect prediction models. IEEE Transactions on Software Engineering, 25:675–689, 1999.
[12] Norman Fenton, Paul Krause, Martin Neil, and Crossoak Lane. A probabilistic model for software defect prediction, 2001.
[13] Norman Fenton, Martin Neil,William Marsh, Peter Hearty, David Marquez, Paul Krause, and Rajat Mishra. Predicting software defects in varying development lifecycles using Bayesian nets. Inf. Softw. Technol., 49(1):32–43, 2007.
[14] Qasem A. Al-Radaideh, Eman Al Nagi "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012
[15] Gaganjot Kaur, Amit Chhabra "Improved J48 Classification Algorithm for the Prediction of Diabetes" International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014
[16] Marcano-Cedeno, Alexis; Andina, Diego, "Data mining for the diagnosis of type 2 diabetes," World Automation Congress (WAC), 2012 , vol., no., pp.1,6, 24-28 June 2012.
[17] Robu, R.; Hora, C., "Medical data mining with extended WEKA," Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on , vol., no., pp.347,350, 13-15 June 2012
[18] Salama, G.I.; Abdelhalim, M.B.; Zeid, M.A., "Experimental comparison of classifiers for breast cancer diagnosis," Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on , vol., no., pp.180,185, 27-29 Nov.,2012.
[19] UM, Ashwinkumar, and Anandakumar KR. "Predicting Early Detection of Cardiac and Diabetes Symptoms using Data Mining Techniques.",IEEE,pp:161-165,2011
[20] Y. Dong, A. K. Milne and B. C. Forster, "Segmentation and classification of vegetated areas using polarimetric SAR image data", IEEE Transactions on Geoscience and Remote Sensing, vol.39, no.2
[21] J. F. Peters, Z. Suraj, S. Shan, S. Ramanna, W. Pedrycz and N. Pizzi, "Classification of meteorological volumetric radar data using rough set methods", Pattern Recognition Letters, vol.24, no.6.
[22] J. Yu, "A support vector clustering-based probabilistic method for unsupervised fault detection and classification of complex chemical processes using unlabeled data", AIChE Journal, vol.59, no.2, (2013), pp.407-419.
[23] P. Kemal, "Data weighting method on the basis of binary encoded output to solve multi-class pattern classification problems", Expert Systems with Applications, vol.40, no.11, (2013), pp.4637-4647.
[24] P. Li, X. Y. Yu, T. T. Bi and J. L. Huang, "Imbalanced data SVM classification method based on cluster boundary sampling and DT-KNN pruning", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol.7, no.2, (2014), pp.61-68.
[25] D. L. Wang and M. Shi, "Density weighted region growing method for imbalanced data SVM classification in under-sampling approaches", Journal of Information and Computational Science, vol.11, no.18, (2014), pp.6673-6680.
[26] Z. B. Liu, Y. Y. Gao and J. Z. Wang, "Automatic classification method of star spectra data based on manifold fuzzy twin support vector machine", Spectroscopy and Spectral Analysis, vol.35, no.1, (2015), pp.263-266.
[27] J. Cao, L. Zhang, B. J. Wang, F. Z. Li and J. W. Yang, "A fast gene selection method for multi-cancer classification using multiple support vector data description", Journal of Biomedical Informatics, vol.53, no.1, (2015), pp.381-389.

BIOGRAPHIES

Akash Sharma currently pursuing M.Tech (CSE) from GIT College Jaipur affilated to Rajasthan Technical University, Kota. He did B.TECH in Computer Science and Engg. from BMIT, Jaipur in 2010. His interested research areas are Data Mining. Computer Networks.

Ms. Nikita Jain obtained B.Tech. Degree in Computer Science and Engineering from UPTech University lucknow in 2005 and also completed M.Tech. in 2014 with same subject from Rajasthan Technical University, Kota. She has published Several research papers in reputed conference & journals.