ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

Classification Algorithms - Support Vector Machine, Back Propagation Neural Network and K-Nearest Neighbor: A Review

K.Pavya[1], Dr.B.Srinivasan[2]

Assistant Professor, Department of Computer Science, Vellalar college for women, Erode, Tamil Nadu, India [1,]

Associate Professor, Department of Computer Science, Gobi arts and science college, Gobichettipalayam, Tamil Nadu, India[2]

ABSTRACT - Data Mining is a pioneering and attractive research area due to its vast application areas and task primitives. In broader point of view, we have reviewed the number of research publications that have been contributed in various internationally reputed journals for the data mining applications and also suggested a possible number of issues in SVM (Support vector Machine), BPNN (Back Propagation Neural Network) and KNN (K-Nearest Neighbor). The main aim of this paper is to extrapolate the various areas of the above three classification algorithms with a basis of understanding the technique and a comprehensive review, with their merits and demerits.

KEYWORDS: Data mining, Classification, Support vector machine (SVM), Back propagation neural network (BPNN) and K-nearest neighbor (KNN).

I. INTRODUCTION

Classification in data mining is a technique based on machine learning algorithms which uses mathematics, statistics, probability distributions and artificial intelligence. It is a data mining function that assigns items in a group to target Categories or classes. The objective of classification is to accurately predict the target class for each case in the data. It is a supervised learning and can be used in designing of models describing data classes, where attribute class is required in the construction of classifier. Classification is useful in categorizing the object that has similar functionality.

The simplest type of classification problem is binary classification. The target attribute has only two possible values in binary classification: for example, high credit rating or low credit rating. Multiclass targets have more than two possible values: for example, low, medium, high, or unknown credit rating. The classification algorithm discovers relationships between the values of the predictors and the values of the target in the training process. Different Classification algorithms use different technique for finding relationships. These relationships are reviewed in a model, which can then be applied to a different dataset in which the class assignments are unknown. Classification models are tested by comparing the predicted values to known target values in a set of test data[1].

This paper is structured as follows: Section 2 explains the three commonly used classification algorithms. Section 3 gives the advantages and disadvantages of them. Section 4 provides a review of literature for the three classification algorithms and Section 5 concludes our work.

II. CLASSIFICATION ALGORITHMS

2.1 *K-Nearest Neighbor Algorithm:*

K- nearest neighbors (KNN), a classification scheme based on the distance measure. The KNN technique assumes that the entire training set includes not only the data in the set but also the preferred classification for each item. When classification is to be made for a new item, the distance to each item in the training set must be determined.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

Only the K closest entries in the training set are considered further. The new item is next placed in the class that contains the most items from this set of K closest items [2].

2.2 ***Back Propagation Neural Network:***

Backpropagation is a training method used for a multi layer neural network. It is also called the generalized delta rule. It is a gradient descent method which minimizes the total squared error of the output computed by the net. Any neural network is expected to respond correctly to the input patterns that are used for training which is termed as memorization and it should respond reasonably to input that is similar to but not the same as the samples used for training which is called generalization.

The training of a neural network by back propagation takes place in three stages

1. Feedforward of the input pattern
2. Calculation and Back propagation of the associated error
3. Adjustments of the weights

After the neural network is trained, the neural network has to compute the feedforward phase only. Even if the training is slow, the trained net can produce its output immediately [3].

2.3 ***Support Vector Machine Algorithm:***

Support Vector Machine (SVM) is based on statistical learning theory and is increasingly becoming useful in data mining. The main idea is to non-linearly map the dataset into a high-dimensional feature space and use a linear discriminator to classify the data. Its success has been demonstrated in the areas of regression, classification and decision tree construction. Unlike other classification techniques, which attempt to minimize error of classification, SVMs incorporate structured risk minimization which minimizes an upper bound on the generalized error. Consider two sets A and B is linearly separable. The idea is to determine from an infinite number of planes correctly separating A and B, the one which will have the smallest generalization error. SVMs select with which minimizes the margin separating the two classes. The margin is defined as the distance between the separating hyperplane to the nearest point of a, plus the distance from the hyperplane to the nearest point in B [2].

III. ADVANTAGES AND DRAWBACKS

3.1 K-Nearest Neighbor Algorithm:

Advantages:

- It is an easy to understand and easy to implement classification technique.
- Robust to noisy training data.
- Training is very fast.
- It is particularly well suited for multimodal classes

Drawbacks:

- It is sensitive to the local structure of the data.
- Being a supervised learning lazy Algorithm i.e., runs slowly.
- Memory limitation
- Computation complexity.

3.2 Back Propagation Neural Network:

Advantages:

- It is simple and easy to program
- It has no parameters to tune(except for the number of input)
- It requires no prior knowledge about the weak learner and so can be flexible.
- Computing time is reduced if the weights chosen are small at the beginning

Drawbacks:

- Back propagation can be sensitive to noisy data and outliers.
- The actual performance of back propagation on a particular problem is clearly dependent on the input data.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

- Slow and inefficient

3.3 Support Vector Machine Algorithm:

Advantages:

- Produce very accurate classifiers.
- Memory intensive.
- Mainly popular in text classification problems where very high dimensional spaces are the norm.
- Less over fitting, robust to noise.

Drawbacks:

- SVM is a binary classifier. To do a multiclass classification, pair-wise classifications can be used (one class against all others, for all classes).
- Computationally expensive, thus runs slow.

IV. LITERATURE REVIEW

Adil Baykasog lu *et. al.*, [4] have presented a comparative analysis of the performances of DE,GA, and FA on both static and dynamic multidimensional knapsack problems. One of the main contributions of their study is the development of FA2, which is designed for a more realistic reflection of the behaviors of fireflies and it requires less computational time. Particularly on stationary environments, FA2 obtains near optimal results with a significantly faster convergence capability. Thus, they assumed that FA2 was found more effective compared to GA, DE and FA for the problems studied.

Jui-Sheng Chou *et. al.*, [5] have discussed some classifiers that can be applied when using CART, QUEST, C5.0, CHAID, and GASVM (a hybrid approach) to predict dispute propensity. In terms of accuracy, GASVM (89.30%) and C5.0 (83.25%) are the two best classification and regression-based models in predicting project disputes. From all the models, GASVM provides the maximum overall performance measurement score (0.871) considering accuracy, precision, sensitivity, and AUC. Notably, with the exception of GASVM, which was developed by the authors and implemented within a mathematical tool, all models are easily executed via open-source or commercial software. Compared to the baseline models (*i.e.*, C5.0, CHAID, CART, and QUEST) and previous work, GASVM provides 5.89–12.95% higher classification accuracy.

Zuriani Mustaffa *et. al.*, [6], has reported empirical results that examine the feasibility of eABC-LSSVM in predicting prices of the time series of interest. The performance of their proposed prediction model was estimated using four statistical metric, namely PA,MAPE, SMAPE and RMSPE and experimented using three different set of data arrangement, in order to choose the best data arrangement for generalization purposes. In addition, the proposed technique also has proven its capability in avoiding premature convergence that finally leads to a good generalization performance.

Youngdae Kim *et. al.*, [7] have proposed an exact indexing solution for the SVM function queries, which is to find top-k results without evaluating the entire database. They first suggested key geometric properties of the kernel space, ranking instability and ordering stability , which is essential for building index in the kernel space. Based on that, they developed an index structure iKernel and processing algorithms and then presented clustering techniques in the kernel space to enhance the pruning effectiveness of the index. According to their experiments, iKernel is highly effective overall producing 1–5% of evaluation ratio on large data sets.

A.D. Dileep *et. al.*, [8] have proposed two novel methods to build a better discriminatory IMK-based SVM classifier by considering a set of virtual feature vectors specific to each class depending on the approaches to multiclass classification using SVMs. They proposed a class-wise IMK based SVM for every class by using components of GMM built for a class and a pair wise IMK based SVM for every pair of classes by using components of GMM built for a pair of classes as the set of virtual feature vectors for that pair of classes in the one-against-one approach to multiclass classification. The performance of the SVM-based classifiers using the proposed class-specific IMKs is considered for

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

speech emotion recognition and speaker identification tasks and compared with that of the SVM-based classifiers using the state-of-the-art dynamic kernels.

Shifei Ding *et. al.*, [9] have enhanced LSPTSVM to nonlinear LSPTSVM(NLSPTSVM)for solving nonlinear classification problems efficiently. Like LSPTSVM, NLSPTSVM requires just the solution of two systems of linear equations in contrast to PTSVM which requires solving two QPPs. They proposed a novel nonlinear recursive algorithm to improve and boost NLSPTSVMs performance. Experimental results shows that NLSPTSVM has good nonlinear classification capability.

Zhong Yin [10] has work with two psycho physiological-data-driven classification frameworks. It is for operator functional states (OFS) assessment in safety-critical human-machine systems with stable generalization ability. They combined the recursive feature elimination (RFE) and least square support vector machine (LSSVM) and used for binary and multiclass feature selection. Feature selection results have discovered that different dimensions of OFS can be characterized by specific set of psycho physiological features. Performance evaluation studies shown that reasonable high and stable classification accuracy of both classification frameworks can be achieved if the recursive feature elimination procedure is correctly implemented and utilized.

Zhen Guo che, Tzu-An Chiang and Zhen Hua Che [11] has done the comparison between genetic algorithm and back propagation learning algorithm over different problems such as Sin function, Iris plant and Diabetes datasets and found that BPLA is faster than GA in terms of training speeds and also in terms of CPU time required, the BPLA yields better results than GA.

Chukwuchekwa Ulumma Joy[12] have studied that using pattern recognition problems, assessments are made based on the effectiveness and efficiency of both back propagation and genetic algorithm, training algorithms on the networks. The back propagation algorithm is found to be better than the genetic algorithm in this instance.

Yeremia, Hendy, et al [13] have discussed that back propagation network algorithm is mingled with genetic algorithm to attain both accuracy and training fastness for recognizing alphabets. The training time needed for back propagation learning phase developed significantly from 03 h, 14 min and 40 sec, a standard back propagation training time, to 02 h 18 min and 1 sec for the optimized back propagation network. A hybrid approach proves to be improving the network performance.

Otair and Salameh[14] studied three algorithms and different versions of backpropagation training for stable learning and robustness to oscillations. The new modification consists of a simple change in the error signal function. These algorithms have been tested on OR problem, encoding problem and character recognition.
Rimer and Martinez [15] have proposed Classification-Based (CB) cost functions that attempted to guide the network directly to correct pattern classification. It reduces average test error. However, these functions only applicable for Classification-based (CB) problems.

Lv andYi[16] Presented an absolute cost function using the Lyapunov method in 2005. This algorithm is more robust and faster for learning than standard BP when target signals include some incorrect data. It also avoids local minima. Samsuddin et al [17] have proposed new modified cost function (MM). It has been proved to improve convergence speed of standard backpropagation. However, for some datasets, the network does not converge.

A study [18] shows the sensitivity, specificity, and accuracy results of KNN in the diagnosis of heart disease patients. The value of K ranged between one and thirteen. The accuracy achieved is between the range of 94% and 97.4 % with different values of K. The value of K equal to 7 achieved the highest accuracy and specificity (97.4% and 99% respectively). This study also indicates that k-NN is one of the most commonly used data mining techniques in classification problems. Because of its ease and relatively high convergence speed make it a popular selection. On the

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

other hand a main drawback of KNN classifiers is the large memory requirement needed to store the whole sample. When the sample is large, response time on a sequential computer is also large.

P. Cunningham et al [19] talk about the most straightforward classifier in the machine learning techniques is the Nearest Neighbor Classifier – classification is achieved by identifying the nearest neighbors to a query example and using those neighbors to determine the class of the query. This approach to classification is of particular importance today because issues of poor run-time performance are not such a problem these days with the computational power that is available. This paper presents an outline of techniques for Nearest Neighbor classification focusing on; mechanisms for assessing similarity (distance), computational issues in identifying nearest neighbors and mechanisms for reducing the dimension of the data.

M. Muja et al [20] discussed for many computer vision and machine learning problems, large training sets are key for good performance. But, the most computationally costly part of many computer vision and machine learning algorithms consists of finding nearest neighbor matches to high dimensional vectors that characterize the training data. They suggest new algorithms for approximate nearest neighbor matching and evaluate and compare it with earlier algorithms. From that, they find two algorithms to be the most efficient, the randomized k-d forest and the priority search k-means tree. They also find a new algorithm for matching binary features by searching multiple hierarchical clustering trees.

N. Kumar et al [21] addresses many computer vision algorithms require searching a set of images for similar patches, which is a very expensive operation. In this work, they compare and evaluate a number of nearest neighbors algorithms for speeding up this task. Since image patches follow very different distributions from the uniform and Gaussian distributions that are typically used to evaluate nearest neighbors methods, they determine the method with the best performance via extensive experimentation on real images. Furthermore, they take advantage of the inherent structure and properties of images to achieve highly efficient implementations of these algorithms. Their results indicate that vantage point trees, which are not well known in the vision community, generally offer the best performance.

M. J. Prerau [22] discussed most current intrusion detection methods cannot process large amounts of audit data for real-time operation. In this work, anomaly network intrusion detection method based on Principal Component Analysis (PCA) for data reduction and Fuzzy Adaptive Resonance Theory (Fuzzy ART) for classifier is presented. Moreover, PCA is applied to reduce the high dimensional data vectors and distance between a vector and its projection onto the subspace reduced is used for anomaly detection. Using a set of benchmark data from KDD (Knowledge Discovery and Data Mining) Competition designed by DARPA for demonstrate to detection intrusions. Experimental results show the proposed model can classify the network connections with satisfying performance.

T. Liu et al [23] addresses about non-approximate acceleration of high-dimensional nonparametric operations such as k nearest neighbor classifiers. They effort to use the fact that even if we want exact answers to nonparametric queries, we generally do not need to explicitly find the data points close to the question, but merely need to reply queries about the properties of that set of data points. This offers a little amount of computational leeway, and investigates how much that leeway can be exploited. But for further clarity, this work concentrates on pure k-NN classification and introduce new ball-tree algorithms that on real-world data sets give accelerations from 2-fold to 100-fold compared against highly optimized traditional ball-tree-based k- NN.

V. CONCLUSION

This paper deals with the three mainly used classification algorithms in data mining. Data mining is a wide area that integrates techniques from various fields including machine learning, artificial intelligence, statistics and pattern recognition, for the analysis of large volumes of data. From this paper we conclude that even though few limitations are

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

in all the three algorithms each will have some merits on their own. Depending upon the classification problem the algorithms can be selected to find out the best results.

REFERENCES

[1] V.Krishnaiah, Dr.G.Narsimha and Dr.N.Subhash Chandra, "Survey of classification techniques in data mining" International journal of Computer science and Engineering, vol.2,(2014),pp.65-74.
[2] K.pavya and Dr,B.Srinivasan, "A Comparative Study on Classification Algorithms in Data Mining", International Journal of Innovative Science, Engineering & Technology, vol.3,(2016),pp.415-418.
[3] NPTEL – Electronics & Communication Engineering – Pattern Recognition
[4] A. Baykasoglu and F. B. Ozsoydan, "An improved firefly algorithm for solving dynamic multidimensional knapsack problems", Expert Systems with Applications, vol. 41, (2014), pp. 3712–3725.
[5] J. S. Chou, M. Y. Cheng, Y. W. Wu and A. D. Pham, "Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification", Expert Systems with Applications, vol. 41, (2014), pp. 3955–3964.
[6] Z. Mustaffa, Y. Yusof and S. S. Kamaruddin, "Enhanced artificial bee colony for training least squares supportvector machines in commodity price forecasting", Journal of Computational Science, vol. 5, no. 2, (2014) March, pp. 196–205.
[7] Y. Kim, I. Ko, W. S. Han and H. Yu, "iKernel: Exact indexing for support vector machines", Information Sciences, vol. 257, (2014), pp. 32–53.
[8] A. D. Dileep and S. C. Chandra, "Class-specific GMM based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines", Speech Communication, vol. 57, (2014), pp. 126–143.
[9] S. Ding and X. Hua, "Recursive least squares projection twin support vector machines for nonlinear classification", Neurocomputing, (2014).
[10] Z. Yin and J. Zhang, "Operator functional state classification using least-square support vector machine basedrecursive feature elimination technique", Computer methods and programs in biomedicine, vol. 113, (2014), pp. 101-115.
[11] Che, Zhen-Guo, Tzu-An Chiang, and Zhen-Hua Che. "Feed-forward neural networks training: a comparison between genetic algorithm and back-propagation learning algorithm." Int J Innov Comput Inf 7 (2011): pp.5839-5850.
[12] Joy, Chukwuchekwa Ulumma. "Comparing the Performance of Backpropagation Algorithm and Genetic Algorithms in Pattern Recognition Problems."International Journal of Computer Information Systems vol. 2, no.5 (2011),pp.7-12.
[13] Yeremia, Hendy, et al. "Genetic Algorithm and Neural Network for Optical Character Recognition." Journal of Computer Sciencevol.9, no.11 (2013), pp. 1435-1442.
[14] Otair, Mohammed A., and Walid A. Salameh. "Efficient training of backpropagation neural networks." Neural Network World vol.16,no.4 (2006).
[15] Rimer, Michael, and Tony Martinez. "CB3: an adaptive error function for backpropagation training." Neural processing letters vol.24,no.1 (2006),pp. 81-92.
[16] Lv, Jiancheng, and Zhang Yi. "An improved backpropagation algorithm using absolute error function." Advances in Neural Networks–ISNN 2005. Springer Berlin Heidelberg, (2005),pp. 585-590.
[17] Shamsuddin, Siti Mariyam, Md Nasir Sulaiman, and Maslina Darus. "An improved error signal for the backpropagation model for classification problems."International Journal of Computer Mathematics vol.76,no.3 (2001),pp. 297-305.
[18] Davis LS, Terdiman J , "The medical data base, chap 4. In: Collen MF, editor". Hospital computer systems. New York: Wiley, (1974), pp. 52–79.
[19] P. Cunningham and S. J. Delany, "k-nearest neighbour classi- fiers," Multiple Classifier Systems, (2007) pp. 1–17,
[20] M. Muja and D. Lowe, "Scalable nearest neighbour algorithms for high dimensional data," IEEE Trans. Pattern Anal. Mach.Intell., vol. 36, no. 11, (2014) pp. 2227–2240.
[21] N. Kumar, L. Zhang, and S. Nayar, "What is a good nearest neighbours algorithm for finding similar patches in images?" in Proc.10th Eur. Conf. Comput. Vis., (2008), pp. 364–378.
[22] M. J. Prerau and E. Eskin, "Unsupervised anomaly detection using an optimized k-nearest neighbors algorithm,"Undergraduate thesis, Columbia Univ., New York, NY, USA, Dec. 2000.
[23] T. Liu, A. W. Moore, and A. Gray, "New algorithms for efficient high-dimensional nonparametric classification," J. Mach.Learning Res., vol. 7,(2006) pp. 1135–1158.