International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Efficient Classification Rules for Complex Data in Decision Making

L Kiran Kumar Reddy, S Phani Kumar

Abstract*: Information mining in enterprise applications facing challenges due to the complex data distribution in large heterogeneous sources. In such scenario, a single approach or method for mining limited the information needs and it also will be a high processing and time consuming. It is necessary to develop an effective mining approach which can be useful for the real time business requirements and decision making tasks. This paper proposed an efficient classification rules generation mechanism for complex data association and information mining using Multi-Features Patterns Combination (MFPC) method. The approach builds a strong association rule between multi features patterns using Feature reduction which will be used for efficient classification for complex data. The approach is evaluated in comparison with the existing feature reduction and classification approaches and measure the classification accuracy to show the flexibility and capability of the proposed mechanism in data classification.

Keywords: Data Mining, Classification, Complex Data, Association Rules, Feature Reduction, Decision Making.

I. INTRODUCTION

A wide variety of fields, data are being collected and accumulated at a impressive pace.There is an urgent need for a new generation of computational approach and tools to assist humans in extracting useful information from the rapidly growing volumes of digital data and vast data storage [7],[8],[9]. It is a great challenge in the information age turning data into information and turning information into knowledge from the anomalies datasets. However, most modern businesses systems carries multiple datasets which have heterogeneous characteristics and also are of larger sized for their daily business analysis. In real time, this distributed data is requires a larger space and much time for processing if multiple sources are joined. Table joining is a mostly adopted mining technique to fetch the patterns from tables which are associated with multiple relations. This technique formulates a composite pattern by extracting the relevant features from individual tables. Hence this composite pattern integrates the nature of multiple features from different tables. This mining technique is more adoptable for the multiple relational databases, especially for datasets which are of smaller size and most of the existing Approaches [1],[6],[11],[12] for classification and pattern generation are less efficient in producing required results in multiple and complex data sources[4]. Clustering and classification based approaches based on association rule mining [3], [20], [22], [23], sequence classification [25] and combing clustering and association rules for mining [2] have a common problem of efficiency and effective selection of discriminative features from large feature data space[17]. The existing methods have more challenges to mine informative and comprehensive knowledge in some complex datasets which are more nearer to the real life decisions [4], [10]. There are so many challenges, for example the existing methods only focused over the discovery of homogeneous features from a data related to single source while it is not effective to mine the patterns which are associated with multiple data source components. Moreover, the mining of such kind of patterns is associated with heavy cost or sometimes it is impossible. This paper propose a Multi-Features Patterns Combining (MFPC) approach for data association and classification rules generation to overcome the tradition approach limitation with the selective approaches on basis of the multi attributes pattern combination. The following paper organized in five sections. In Section-2 we present an insight on the background works, Section 3, discuss the proposed classification rules for complex data association, Section 4, presents the Experiment Evaluation and Section-5 presents the conclusion and future work.

II. BACKGROUND STUDY

The informative data discovering is having a crucial role in following organizations telecom, large scale industry, heterogeneous and distributed data sources inserting information about business transactions. These organizations will focus on discovering the data which in turn is informative and helpful to produce some business logic rules with the help of stable with multiple objects. But the data discovered, is from multiple sources and providing the informative data is difficult, as the organizations need to consider the different sources and different features to providing the informative data. The most of the existing data mining methods not concrete in discovering patterns on complex data.

Revised Manuscript Received on March 30, 2020.
***** Correspondence Author
L Kiran Kumar Reddy*, Research Scholar, Dept. of CSE, GITAM University, Hyderabad, India
Dr. S. Phani Kumar, Professor & HOD, Dept. of CSE, GITAM University, Hyderabad, India

Direct mining for discriminative patterns has been highlighted, such as firstly Harmony [14] and model-based search tree [13].

The Harmony approach works on classification rules to directly mine the final set. It uses an instance centric rule generation approach in the sense that it can assure for each training instance, one of the highest confidence rules covering this instance is included in the result set, which helps a lot in achieving high classification accuracy. Model-based search tree builds a decision tree that partitions the data onto different nodes. Then at each node, it directly discovers a discriminative pattern to further divide its examples into purer sub sets. This approach able to examine patterns with extremely minimum support and the discovered features might be more accurate on some of the most difficult graph. But, the minimum support of some discriminative patterns can be extremely low support patterns and attains huge memory consumption. Multi-relational classification [15] has been a widely accomplished technique is several applications like Geographical, medical research, decision making in financial streams, etc. However, most of the classification models are worked on flat or single data relations. Generally, it is much difficult to translate multiple relations into a single flat relation. Even though if it is accomplished, the method faces either a huge significant information loss or an undesirable huge "universal relation". The approaches based on relational mining like Cross Mine [15] have gained an efficient accurate performance in the classification of multi-relational data. However, they have gained a very less scalability with respect to attribute count in databases and also with relations count. Generally, the dataset is associated with several association rules, which are the basic class of patterns [16], [18], [21]. Completeness is the key factor which defines the strength a mining based on association rules. It discovers all the associations with which the data is associated such that the user will satisfy with required minimum confidence constraints and minimum support. However, the major drawback of this method is discovery of a larger number of associations, especially for the dataset which have highly correlated attributes. For a user, the larger associations count makes much difficult, sometimes impossible, to identify and analyse the required ones. Instead of mining the combined patterns directly [5], processing the discovered patterns to post mining process makes the patterns more actionable. For instance, in the multi feature combined mining method, the feature from multiple datasets is considered directly during the generation of more significant patterns [1],[3],[9],[23].

This paper aim to enhance the methods for multi relational data mining approaches through a multi-features patterns combining approach for mining complex knowledge in complex data for various decision making needs. The enhance in classification rules are compiles in combining the multi features pattern association with multiple data sets.

III. PROPOSED MULTI-FEATURES PATTERNS COMBINING APPROACH

The integration of multiple data mining methods is widely used to mining for more informative knowledge, such as associative classification [**27**]for efficient data analysis. Multi-features Patterns Combining (MFPC) enhance these approach to propose a new classification and association rules mining for the complex data. The proposed approach defines the mechanism for building efficient classification rules using multiple features patterns for accurate classification.

Associative classification is a novel and powerful method originating from association rule mining for building associate classifiers[22],[24]. Mostly, small number of high-quality association rules were used in the prediction, decision and classification .MFPC combined highly associated patterns obtained from multiple heterogeneous item sets of different datasets to form a combined rules patterns which will be highly efficient for data classifiers and useful for decision making. MFPC implements two mechanisms to build efficient rules. It initially find the highly associated features through feature reduction mechanism and for the reduced features patterns are generated which are combines to form the efficient classification rules.

3.1 Feature Reduction for Multi-Feature Pattern Combination

Feature reduction is an effective mechanism to find highly interesting features required for the classification. Some features might have abundant redundancy of data due to the inconsistency of resource generated. We measures each features association using covariance deviation (*CD*) to find the highly impacting features.

For a give two features we compute the strong implies using a probability and statically covariance deviation between two or more features. Let's assume, *X* and *Y* are the two features having a unique set of *k* values as $\{(x_1,y_1), \ldots, (x_k,y_k)\}$, and the entropy of these values of *X* and *Y* is computed using the equation-(1) and (2) as,

$$H(X) = \bar{X} = \frac{\sum_{v=1}^{k} x_v}{k} \quad (1)$$

and,

$$H(Y) = \bar{Y} = \frac{\sum_{v=1}^{k} y_v}{k} \quad (2)$$

Based on the computed entropy of the feature we compute each features CD variance using equation-(3) as,

$$CD_A = \sum_{v=i+1}^{k} H(X) - H(Y_v) \quad (3)$$

Utilizing the features *CD* values we creates sets of features as, *F* which are >=1. The features which CD value is < 1 are considered as low variance and less impact on classification and the features which are >=1 are considered as high variance and have impacts on classification. Now, using only *F* reduced features we build the classification rules combining the patterns of each features.

3.2Classification Rules using Multi-Features Pattern Combination

Association-rule-mining-based classification [19],[26] is a rule based classifier. In this process, the system learns the rules from a set of instances of training data which are assigned with class labels and uses the trained rule to classify the new incoming instance of data. In this method, the data imperfections are accommodated by a probabilistic relational model in which the attributes are represented through probabilistic functions.

This approach can remove the imperfections in data more effectively. However, the main task is to reduce the computational burden which is associated with the extraction of the items patterns and rules from relational databases. So, to build a low computational overhead and efficient classification rules, we utilize the reduced features set, *F* obtained using equation-(3) and its unique item sets to generate each feature pattern to perform MFPC as discussed below.

Let's assume a set a datasets as, D_n have a *F* reduced features sets which builds a *P* patterns on classification using features values. To generate the individual patterns for each features we utilize a association rule mining method using equation-(4) as,

$$P_n = R(F_k) \quad (4)$$

where, $P_n$is the extracted pattern of each features and n=*1, ... , N,R* is the data mining method used for item set extraction and $F_k$is the features value from *k=1, ... , K*.

The obtain patterns, P_n of each features, F_k, will be merge to generate a combined pattern as P_k for each dataset of D_n.

$$P_k = C_F(P_n) \quad (5)$$

Using, the equation (4) and (5), a new classification rule will be formed for all dataset, D_n in combine as,

.

$$P_n = R\ (F_1 \wedge \dots \dots \wedge F_k) \longrightarrow A_k \quad (6)$$

$$P := C\ (P_1 \wedge \dots \dots \wedge P_n) \longrightarrow Q \quad (7)$$

where, A_k is refer as associated patterns and *Q* is refer as qualified patterns for the classification rules. If the features are associated with the target data item sets then the data record can be consider as qualified for the decision making.

To perform a real-time complex data association, let's assume a set of test data as Z_k having a reducedF_x features. A initial features pattern extraction using equation-(4) generates a P_n patterns which will be combined using equation-(5) to get the combined pattern of the input data records as,

$$P_n = R\ (F_1, \dots \dots, F_x) \quad (8)$$

$$P_k = C_F(P_n) \quad (9)$$

Now, the obtained combined pattern P_k will be evaluated to find the correlation of the pattern obtain through the traditional measures as, support, confidence and lift as defined below for each item sets as *E* of the datasets,

$$Support = Prob(E \wedge Q) \quad (10)$$

$$Confidence = Prob(E \wedge Q)/(Prob(E)) \quad (11)$$

$$lift = Prob(E \wedge Q)/(Prob(E) * Prob(Q)) \quad (12)$$

If the lift resulting of equation -(12) is less than one, then the data record is negatively correlated with the classified rules, *Q* patterns, and if it is greater or equal to one then we can say it is positively correlated the rules and qualifies the decision requirements. Mostly multi feature combined mining is used to grouping the various datasets having similar features and making as the cluster patterns and these patterns are generated by applying the association rule algorithm.

IV. EXPERIMENTAL EVALUATION

4.1 Datasets

To perform an experimental evaluation we used the CoIL 2000 dataset [28]which contains information on customers of an insurance company. The dataset consist of 86 features and demographic data of the customer derived from different area codes. The features describes 43 demographic and 43 insurance policies variables. It has total 9822 data records in which5822datasets are descriptions of customers for training set with a class label yes or no for the policy and a 4000 test set data of the customers. The datasets is evaluated to classify the customers who are interested in buying insurance policy using the proposed MFPC method.

4.2Evaluation

To perform the evaluation analysis we implement the feature reduction method using java and performance evaluation is measured using Weka-3.6 Tool. For the evaluation of the feature reduction method we compare them with the existing feature reduction technique such as Information gain and Gain Ratio. The comparison result are present in Table-1.

Table-1: Feature Reduced Sets using Trained Data

FeaturesReduction Methods	Demographic Features Selected (1-43)	Insurance policies Features Selected (44-86)	Total Features Count
Information Gain (*IG*)	36	33	68
Gain Ratio (*GR*)	34	36	70
Proposed FR Method	26	28	54

Based on the features selected we implemented the MFPC mechanism to build the classification rules, and the obtained classification rules are evaluated using Weka Tool in compare with Naive Bayes(NB) and J48classifiers with IG, GR and proposed FR method features selected. To measure the classifier performance we measure the classifier accuracy and error measures as shown in Table-2 and 3 below.

Table-2: Classifiers Accuracy Comparison

Classifiers	Correctly Classified	Incorrectly Classified	Classifier Accuracy
NB+IG	2860	1140	71.5
NB+GR	3108	892	77.7
J48+IG	2218	1782	55.45
J48+GR	2671	1329	66.775
MFPC	3695	320	92.375

Table-3: Classifiers Error Rates Measures Comparison

Classifiers	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root relative Squared Error	Kappa Statistic
NB+IG	0.2141	0.4325	48.321	59.1168	0.7871
NB+GR	0.1635	0.3612	37.236	40.278	0.8141
J48+IG	0.2954	0.5128	52.447	82.879	0.6852
J48+GR	0.1984	0.3951	41.472	68.693	0.7156
MFPC	0.0901	0.1207	21.439	24.421	0.9614

The obtained results of the classifiers are compared and presented in Figure-1, 2 and 3 related to classifier accuracy, relative error rate and root mean error and kappa statistics.

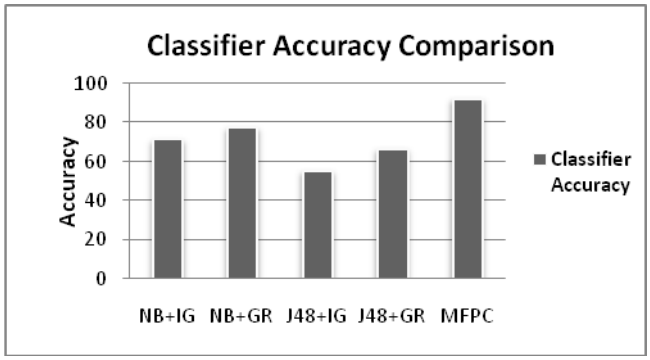

Figure-1: Classifier Accuracy Comparison

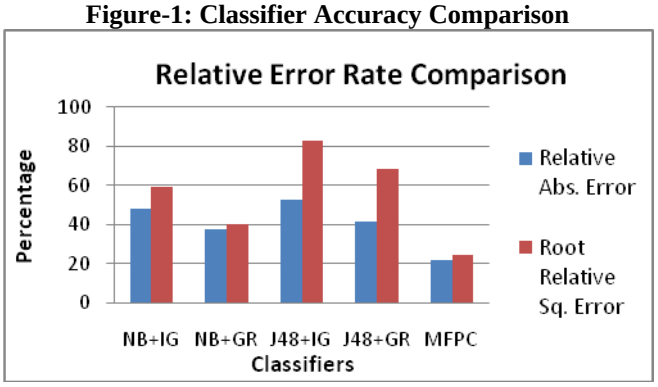

Figure-2: Relative Error Rate Comparison

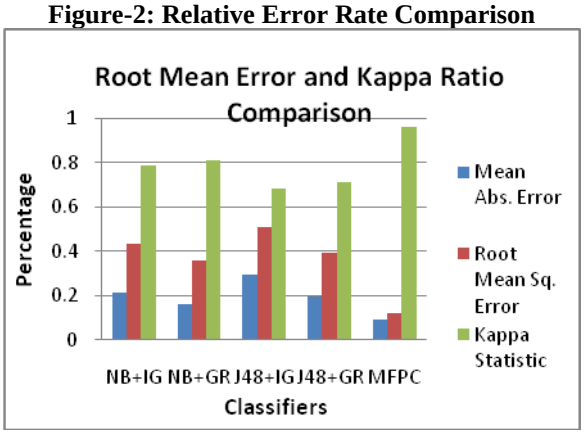

Figure-3: Root Mean Error and Kappa Ratio Rate Comparison

The comparison results in Figure-1,2 and 3 shows that MFPC based classification approach achieves an average of 20% higher accuracy with lower error rate and improvised kappa statistic rate in compare to the existing classifier. It is very important to select accurate features in complex data for the classification to meet the accuracy as high number of feature can deviate the predication cause high number of error. MFPC accurate features selection minimizes the false predication which improvises the accuracy and minimizes the error.

Table-4: MFPC Accuracy varying minimum Support Comparison

Support (%)	Correctly Classified	Incorrectly Classified	Classifier Accuracy
2	3695	305	92.375
4	3491	509	87.275
8	3149	851	78.725
10	3052	948	76.3
15	2893	1107	72.325

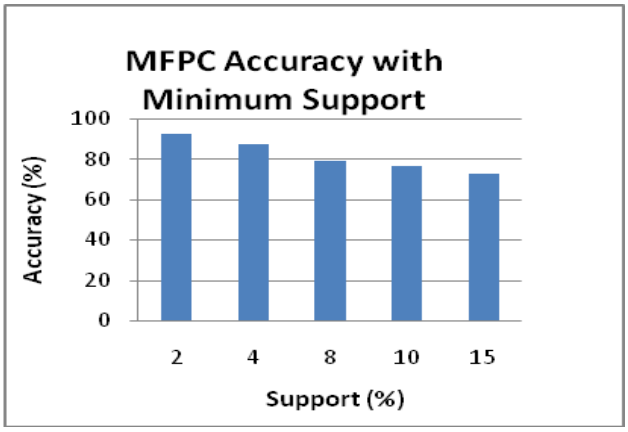

Figure-4: MFPC Accuracy with varying Minimum Support

Figure-4 shows the effects of minimum support threshold varying with a wide range on the MFPC classification accuracy on the datasets. It shows high accuracy of 92% with low-support value as 2% threshold and with minimum support variance a 2-3 % accuracy drop is observed as, increase in support threshold allows relevance patterns selection only for combination to build the classification rules. The interpretation of all the results concludes that the MFPC approach can be comprehensible and useful for various computational learning and analysis in different domains.

V. CONCLUSION

Classification rules using combined feature pattern mining provides a general mechanism for discovering more informative knowledge in complex data. Typical challenges such as mining heterogeneous data sources can benefit from combined pattern mining. The proposed MFPC based classification rules for complex data using association approach is to overcome the tradition table joining. The approach present a feature reduction using covariance deviation makes us to select precise feature and build efficient pattern for combination and classification rules.

The experiment evaluation shows an improvisation in accuracy and low error rate in comparison with existing classifier and feature selection approaches. The improvisation in the classification will be effective in developing combined mining methods to handle the multiple sources data, especially which is available in banking, insurance, and industry projects of government to take critical decisions.

REFERENCES

1. L. Cagliero and P. Garza, "Infrequent Weighted Itemset Mining Using Frequent Pattern Growth", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 4, April 2014.
2. Zechao Li, Jing Liu, Yi Yang, Xiaofang Zhou,Hanqing Lu, "Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 9, September 2014.
3. C Lucchese, S Orlando and R Perego, "A Unifying Framework for Mining Approximate Top-k Binary Patterns", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 12, December 2014.

4. B. Zhang and M. Becker, "Mining complex feature correlations from software product line configurations", In Proceedings of the Seventh International Workshop on Variability Modelling of Software-intensive Systems,VaMoS, ACM, 2013.
5. Longbing Cao, "Combined Mining: Analyzing Object and Pattern Relations for Discovering Actionable Complex Patterns", In Advanced Analytics Institute, University of Tech. Sydney, Australia, Dec 2012.
6. Asif, Muhammad,Ahmed Jamil, "Analysis of Effectiveness of Apriori and Frequent Pattern Tree Algorithm in Software Engineering Data Mining", IEEE Intelligent Systems Modelling and Simulation (ISMS), 6th International Conference on Kuala Lumpur, Malaysia, DOI: 10.1109/ISMS, Feb. 2015.
7. M Tao, F Zhou, Yan Liu and Z Zhang, "Tensorial Independent Component Analysis-Based Feature Extraction for Polarimetric SAR Data Classification", IEEE Transactions On Geoscience And Remote Sensing, Vol. 53, No. 5, May 2015.
8. G J. Simon, P J. Caraballo, T M. Therneau, Steven S. Cha,M. R Castro, and Peter W. Li "Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 1, January 2015.
9. E Baralis, L Cagliero, and P Garza, "EnBay: A Novel Pattern-Based Bayesian Classifier", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 12, December 2013
10. N Hariri, C Castro-Herrera, M Mirakhorli,J Cleland-Huang, B Mobasher, "Supporting Domain Analysis through Mining and Recommending Features from Online Product Listings", IEEE Transactions On Software Engineering, Vol. 39, No. 12, December 2013.
11. E. Baralis, S. Chiusano, and P. Garza, "A Lazy Approach to Associative Classification", IEEE Transactions On Knowledge And Data Engineering, vol. 20, no. 2, pp. 156-171, Feb. 2008.
12. K. Sun and F. Bai, "Mining Weighted Association Rules Without Pre-assigned Weights", IEEE Transactions On Knowledge and Data Engineering, vol. 20, no. 4, pp. 489-495, Apr. 2008.
13. W. Fan, K. Zhang, J. Gao, X. Yan, J. Han, P. Yu, O. Verscheure, "Direct mining of discriminative and essential graphical and itemset features via model-based search tree," in Proc. KDD, pp. 230-238, 2008.
14. J. Wang and G. Karypis, "HARMONY: Efficiently mining the best rules for classification," in Proc. SDM, pp. 205-216, 2005
15. X. Yin, J. Han, J. Yang, and P. S. Yu, "Efficient classification across multiple database relations: A CrossMine approach," Transactions On Knowledge and Data Engineering, vol. 18, no. 6, pp. 770-783, Jun. 2006.
16. Mingzhu Zhang, Changzheng He, "Survey on Association Rules Mining Algorithms", In Advancing Computing, Communication, Control and Management Lecture Notes in Electrical Engineering Volume 56, pp 111-118, 2010.
17. Q. Cheng, H. Zhou, and J. Cheng, "The Fisher-Markov selector: Fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data", IEEE Transactions Pattern Anal. Mach. Intell., vol. 33, no. 6, pp. 1217–1233, Jun. 2011.
18. Yazdi A., Kahani, M. "A novel model for mining association rules from semantic web data", IEEE Intelligent Systems (ICIS), 2014 Iranian Conference on Bam, 978-1-4799-3350-1, Feb. 2014.
19. H. C. Peng, F. H. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundanct",IEEE Transactions Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
20. H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering",IEEE Transactions On Knowledge And Data Engineering, vol. 17, no. 4, pp. 491–502, Apr. 2005.
21. R. Agrawal, T. Imielinski, and Swami, "Mining Association Rules between Sets of Items in Large Databases", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '93), pp. 207-216, 1993.
22. Y. Sun, Y.Wang, and A.Wong, "Boosting an associative classifier," IEEE Transactions On Knowledge and Data Engineering, vol. 18, no. 7, pp. 988-992, Jul. 2006.
23. J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 1-12, 2000.
24. D. Meretakis and B. Wuthrich, "Extending Naive Bayes Classifiers Using Long Itemsets", Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '99), pp. 165-174, 1999.
25. C.K.-S. Leung, C.L. Carmichael, and B. Hao, "Efficient Mining of Frequent Patterns from Uncertain Data", Proc. Seventh IEEE Int'l Conf. Data Mining Workshops (ICDMW '07), pp. 489-494, 2007.
26. F. Tao, F. Murtagh, and M. Farid, "Weighted Association Rule Mining Using Weighted Support and Significance Framework", Proc. 9th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '03), pp. 661-666, 2003.
27. K. Hewawasam, K. Premaratne, and M.L. Shyu, "Rule mining and classification in a situation assessment application: A belief-theoretic approach for handling data imperfections," IEEE Transactions On Syst.,vol. 37, no. 6, pp. 1446-1459, Dec. 2007.
28. Dataset: A Bias-Variance Analysis of a Real World Learning Problem: The CoIL Challenge 2000, Machine Learning, 57, 177–195, 2004, http://www.inf.ed.ac.uk/.

AUTHORS PROFILE

L Kiran Kumar Reddy, His area of research is Data Mining. He is a research scholar pursuing his Ph.D (CSE) from GITAM University under the Guidance of Dr.S.Phani Kumar. He is a life member of ISTE. He has published 2 Scopus journals and contributed 4 contributions towards conferences and journals.

S Phani Kumar, He has published above 20 Scopus/SCI/Web of Science journals, guiding about 15 research scholars and 1 awarded till now. He is a Reviewer for Springer Journal, IGI Global Journal etc. His areas of interest are Data Mining, Machine Learning, Software Quality Assurance, Wireless Sensor Networks, Image Processing etc. he is a member of Indian He is member of Indian Science Congress Association, Indian Society for Technical Education, Computer Society of India.