# A New Approach for Classification Algorithm in Data Mining

M. THILLAIKARASI
*Assistant professor, Department of Computer Science and Engineering, Annamalai University*

***Abstract- Data mining is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, data mining algorithms, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue. I consider classification techniques that are based on statistical and AI techniques to perform controlled experiment data characteristics are systematically altered to introduce imperfections such as nonlinearity, unequal covariance. Two machine learning algorithms Naive Bayes and Support Vector Machine (SVM) is used to build models for the automatic classification of the tweets, and these models were evaluated across the metrics of accuracy, precision, recall, area under curve and measure. The results reveal that the proposed sampling strategy makes more judicious use of data points by selecting locations that clarify high level structures in data, rather than choosing points that merely improve quality of function approximation.***

***Indexed Terms- Data Mining, Knowledge discovery data base, Decision tree, Extreme learning machine, Support Vector Machine.***

## I. INTRODUCTION

Using disasters and emergencies, micro blogs, have been used by people whether from the private or public sector, local or international community, as a medium to broadcast their messages [1]. A configuration comprises many adjustable parameters, and the goal of wireless system characterization is to assess the relationship between these parameters and performance metrics measure of the number of bits transmitted in error using the system [2]. Products of sensor data to a website where it can be browsed by interested users and perhaps downloaded for later analysis [3]. One task is considered that of computing the parameter setting which maximizes the likelihood of all the observations given the models [4]. The spatiotemporal domain, timely identification of such patterns can allow for effective interventions to detecting anomalous increases in upload deaths can enable health care workers to effectively target overdose prevention programs [5]. Internet as a communication and transaction channel provides a means to implement many new enabling technologies such as collaborative filtering and recommender system [6]. Tweets highly vary in terms of subject and content and the influx of tweets particularly in the event of a disaster may be overwhelming. It consists of socio-behaviors that include intensified information search and information contagion [7]. The improved rotation forest based on heterogeneous classifiers to classify gene expression profile to decrease the genes are ranked by using relieff algorithm and the top-ranked genes are selected to build new training subset [8].

## II. LITERATURE REVIEW

There are several researches on text mining for classification and prediction on various domains such as in the medical, business, crime investigation, e-mail detection. The following works are focused on the classification of tweets and the comparison of classifying algorithms [9]. The compared Naive Bayes, SVM and K-nearest Neighbor as implemented in Rapid miner to classify sentiments of tweets as positive negative or neutral using a dataset on general topics such as education, sports and political news [10]. Navigating the mapping from field to abstract description through multiple layers rather than in one giant step allows the construction of modular data mining programs with manageable pieces that can use similar processing techniques at different levels of abstraction [11]. Knowledge discovery and data

mining (KDD) is a thriving sub-discipline of computer science that aims to extract interesting and actionable patterns from multi-dimensional datasets. KDD techniques are typically aimed unstructured discovery tasks, such as finding all patterns of a certain class of phenomenon from a given dataset rather than an easily recognizable one [12]. I am interested in active learning algorithms which use information about each observable function to learn some composite target function is propose a heuristic for actively learning level sets of composite functions of sums for continuous valued input spaces [13]. This heuristic performs the level set find task more efficiently than both random and sequential sampling of the constituent functions using state of the art heuristics [14].

## III. SYSTEM MODEL

It is imperative interleave data collection and data mining and focus sampling at only those locations that maximize well defined notions of relevance and utility. Importantly, we will not need to sample the entire configuration space, only enough so as to identify a region with acceptable confidence. Active data selection has been investigated in a variety of contexts [15]. A sampling strategy typically embodies a human assessment of where might be a good location to collect data. The goal will be to extract the underlying spatiotemporal coherences which are embedded both within a single Earth Cube dataset and across multiple datasets [16]. Additionally, an unconstrained search may return an unrelated set of points, reducing interpretability and increasing the potential for over fitting. The procedure constructs a discriminate function by maximizing the ratio of between groups and within group's variances [17]. The performance results data were tested for normality prior to the computation of the significant differences between the Naive Bayes and SVM [18].

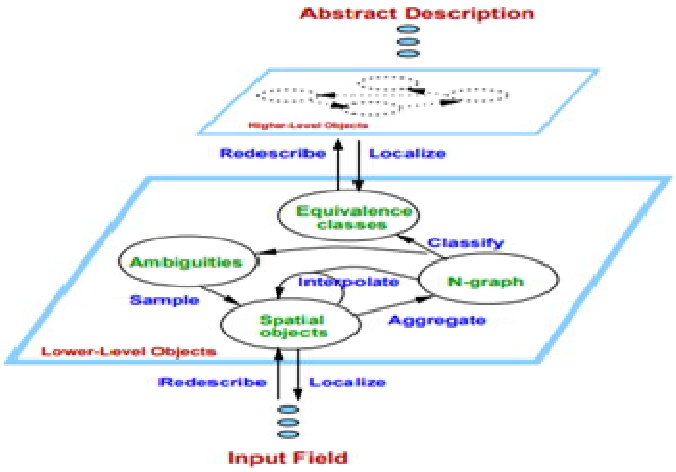



Figure 1: Multi-layer Spatial Aggregates Computations.

## IV. PROPOSED ALGORITHMS

Naive Bayes and Support Vector Machine is most commonly used machine learning algorithms for classification. Naive Bayes classifier is robust and has a good performance in several real-world classification tasks [19]. A Naive Bayes classifier is new probabilistic classifier based on Bayes' theorem with strong independence method [20]. In simple terms Naive Bayes classifier is presence of a particular feature of a class is unrelated to the presence of any other feature. Normality testing is used to determine if the data is normally distributed and consequently whether to use parametric and nonparametric testing of significant difference data is normally distributed, parametric tests are utilized else otherwise [21].

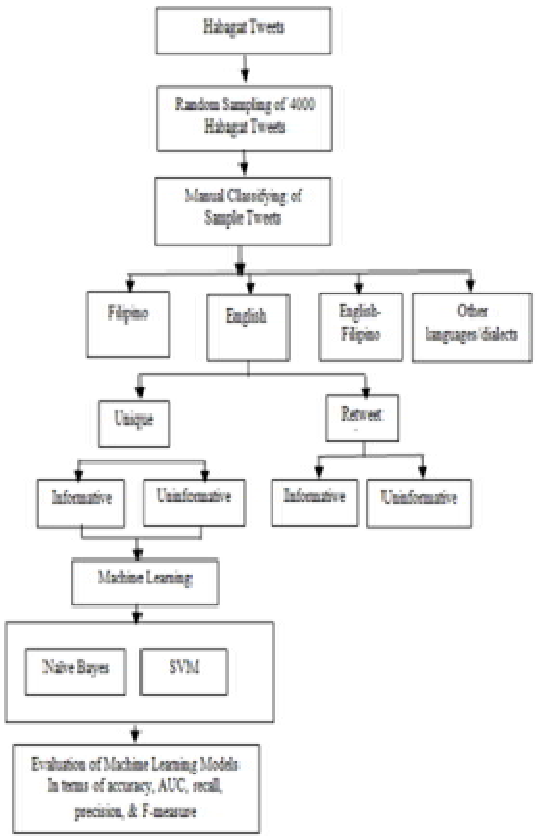



Fig. 2 Methodology Structure

## V. CLASSIFICATION ALGORITHMS

### A. Decision tree (DT)

Decision tree is an effective graphical method to process of classifying or finding new object. It is clear how decisions can be made and models are automatically built from tag samples. Decision trees is constructed using a bottom-up recursive approach which the internal nodes in the tree graph of the decision tree represent the tests on the attributes and the branches represent the outputs of the tests and each leaf node represents a different class [22].

### B. Support Vector Machine (SVM)

Support Vector Machine (SVM) use speed machine learning algorithm is structural problems to minimization for resolving high dimensions, small samples and nonlinear problems. The main idea of SVM is original feature space is mapped onto a high-dimension space by using an appropriate nonlinear function based on Mercer kernel theorem and the original nonlinear classification is converted to a linear classification problem in this high-dimension space and then the optimal hyper plane is found to separate the samples in new feature space [23].

### C. Extreme Learning Machine (ELM)

Extreme learning machine (ELM), proposed is new neural network learning algorithm. ELM has faster learning speed and higher generalization performance because it uses single hidden layer feed forward neural network to reduce the learning time of the algorithm and is widely used in regression and classification problems [24].

- Extreme Learning Machine Algorithm

Input: Training set T = n m n m Q n Q m, activation function (,) t ti g w b x, the number of hidden nodes L
Output: The weight vector connecting hidden nodes and output node β . i.e Y=α+βX
Step 1: Determine the structure of ELM network is practical problem that is the number of nodes of every layers;
Step 2: Randomly generate the weight vector connecting hidden nodes and input nodes wt and the thresholds of hidden nodes t b;
Step 3: Find hidden layer output matrix H according to formula (3);
Step 4: Find weight vector connecting hidden nodes and output nodes β according to formula (4);

## VI. SUPERVISED LEARNING

Supervised learning is a training in which the class attribute values for the dataset are known (labeled data) before running the algorithm [25].

Supervised learning builds a model that maps x to y where x is a vector and y is the class attribute. A model is supervised learning algorithm is run on a training set which maps the feature values (x) to the class attribute values (y). In the context of this study, x = vector of features and y ∈{informative, uninformative}. The dataset is randomly split into 10 mutually exclusive subsets (DS1, DS2...DS10) of approximately equal sizes and with proportional representation of the tweet classes.

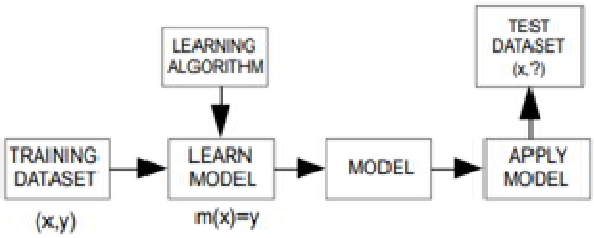



Fig. 2 Supervised Learning

## VII. EXPERIMENT RESULTS

I now present empirical results demonstrating the effectiveness of our active mining strategy on both synthetic and real datasets. The algorithm systematically distorts a convex quadratic function with cubic polynomials to yield continuously differentiable and twice continuously differentiable (D2-type) functions over the closed interval. The first three target functions considered were sums of two observable functions. Mustafa and Elberrichi compared SVM and Naive Bayes in sentiment analysis on Twitter by applying semantics and WorldNet. SVM ranked first with a Measure of 90.75%. New study used SVM, Naive Bayes and MaxEnt for sentiment analysis taking into consideration unigrams, bigrams and emoticons. Experimental results to demonstrate SVM has outperformed the other classifiers. The variations are represented in the graph by means of directions as East, West, and North.

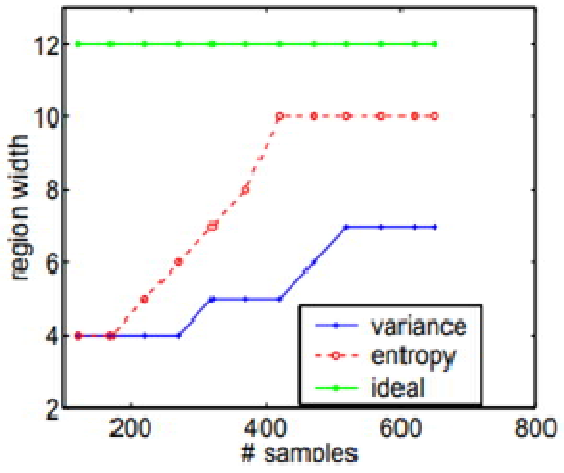
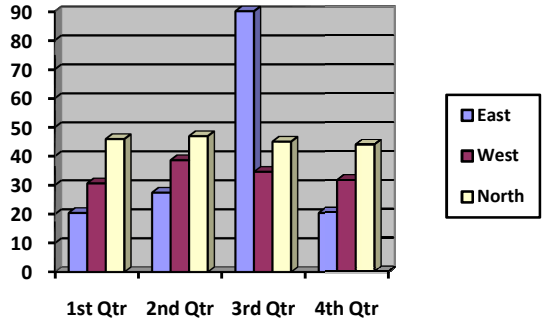
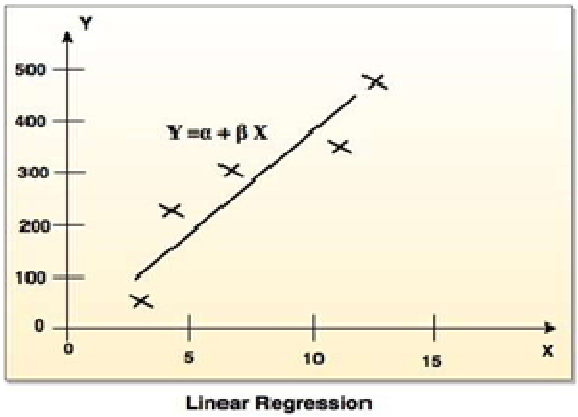



Figure 3: Performance of active mining strategies




## VIII. CONCLUSION AND FUTURE WORK

Data mining offers promising ways to uncover hidden patterns within large amounts of data. These hidden patterns can potentially be used to predict future behavior. The availability of new data mining algorithms, however, should be met with caution. First of all, these techniques are only as good as the data that has been collected. Good data is the first requirement for good data exploration. Assuming good data is available, the next step is to choose the most appropriate technique to mine the data. In this paper, I present the basic classification techniques. Several major kinds of classification method including decision tree, Support Vector Machine, Extreme learning machine. The classification rates due to biases substantially high in the presence of even a single bias. In general more than one method seems to be appropriate candidates based on the type of bias in the data. This evidently implies the informative tweets contain vital and urgent information is provided significantly needed information for situational awareness of the public

and disaster response units. I described and showed how several different heuristics for choosing experiments from a set of candidates perform on synthetic target functions. In this algorithm extreme learning machine, support vector machine and decision tree is used to train multiple heterogeneous base classifiers in ensemble and it can increase diversity among base classifiers to improve ensemble results. Further this algorithm has higher classification and better stability than rotation forest algorithm and it is effective for classifying gene expression profile.

REFERENCES

[1] M. Gaviano, D.E. Kvasov, D. Lera, and Y.D. Sergeyev. Algorithm 829: Software for Generation of Classes of Test Functions with Known Local and Global Minima for Global Optimization. ACM Transactions on Mathematical Software, Vol. 29(4): pages 469–480, Dec 2003.

[2] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, 2000.

[3] C. Bailey-Kellogg, F. Zhao, and K. Yip. Spatial Aggregation: Language and Applications. In Proc. AAAI, pages 517–522, 1996.

[4] K. Brinker. Incorporating Diversity in Active Learning with Support Vector Machines. In Proceedings of the Twentieth International Conference on Machine Learning (ICML'03), pages 59–66, 2003.

[5] D.A. Cohn, Z. Ghahramani, and M.I. Jordan. Active Learning with Statistical Models. Journal of Artificial Intelligence Research, Vol. 4: pages 129–145, 1996.

[6] D. Cornford, I.T. Nabney, and C.K.I. Williams. Adding Constrained Discontinuities to Gaussian Process Models of Wind Fields. In Proceedings of NIPS, pages 861–867, 1998.

[7] C. Bailey-Kellogg and F. Zhao. Influence-Based Model Decomposition for Reasoning about Spatially Distributed Physical Systems. Artificial Intelligence, Vol. 130(2): pages 125–166, 2001

[8] I.T. Nabney. Netlab: Algorithms for Pattern Recognition. Springer-Verlag, 2002.

[9] R.G. Easterling. Comment on 'Design and Analysis of Computer Experiments'. Statistical Science, 4(4):425– 427, 1989.

[10] J. Garcke and M. Griebel. Data Mining with Sparse Grids using Simplicial Basis Functions. In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 87–96, 2001.

[11] C.Bailey-Kellogg and N. Ramakrishnan. AmbiguityDirected Sampling for Qualitative Analysis of Sparse Data from Spatially Distributed Physical Systems. In Proc. IJCAI, pages 43–50, 2001.

[12] M. Gaviano, D.E. Kvasov, D. Lera, and Y.D. Sergeyev. Algorithm 829: Software for Generation of Classes of Test Functions with Known Local and Global Minima for Global Optimization. ACM Transactions on Mathematical Software, Vol. 29(4): pages 469–480, Dec 2003.

[13] X. Huang and F. Zhao. Relation-Based Aggregation: Finding Objects in Large Spatial Datasets. In Proceedings of the 3rd International Symposium on Intelligent Data Analysis, 1999.

[14] J.-N. Hwang, J.J. Choi, S. Oh, and R.J. Marks II. Query-based Learning Applied to Partially Trained Multilayer Perceptrons. IEEE Transactions on Neural Networks, Vol. 2(1): pages 131–136, 1991.

[15] A.G. Journel and C.J. Huijbregts. Mining Geostatistics. Academic Press, New York, 1992.

[16] S. Tong and D. Koller. Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research, Vol. 2: pages 45–66, 2001.

[17] J. Koehler and A. Owen. Computer Experiments. In S. Ghosh and C. Rao, editors, Handbook of Statistics: Design and Analysis of Experiments, pages 261–308. North Holland, 1996.

[18] D.J. MacKay. Information-Based Objective Functions for Active Data Selection. Neural Computation, Vol. 4(4): pages 590–604, 1992.

[19] C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. Bayesian Prediction of Deterministic

Functions, with Applications to the Design and Analysis of Computer Experiments. J. Amer. Stat. Assoc., Vol. 86: pages 953– 963, 1991.

[20] R.M. Neal. Monte Carlo Implementations of Gaussian Process Models for Bayesian Regression and Classification. Technical Report 9702, Department of Statistics, University of Toronto, Jan 1997.

[21] R.T. Ng and J. Han. CLARANS: A Method for Clustering Objects for Spatial Data Mining. IEEE Transactions on Knowledge and Data Engineering, Vol. 14(5): pages 1003–1016, 2002.

[22] I. Ord´onez ˜ and F. Zhao. STA: Spatio-Temporal Aggregation with Applications to Analysis of DiffusionReaction Phenomena. In Proc. AAAI, pages 517–523, 2000.

[23] N. Ramakrishnan and C. Bailey-Kellogg. Sampling Strategies for Mining in Data-Scarce Domains. IEEE/AIP CiSE, Vol. 4(4): pages 31–43, 2002.

[24] N. Ramakrishnan and C. Bailey-Kellogg. Gaussian Process Models of Spatial Aggregation Algorithms. In Proc. IJCAI, pages 1045–1051, 2003.

[25] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and Analysis of Computer Experiments. Statistical Science, Vol. 4(4): pages 409–435, 1989.

[26] G. Karypis, E.-H. Han, and V. Kumar. Chameleon: Hierarchical Clustering using Dynamic Modeling. IEEE Computer, Vol. 32(8): pages 68–75, 1999.