ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(27), DOI: 10.17485/ijst/2016/v9i27/92383, July 2016* ISSN (Online) : 0974-5645

Agriculture Yield Analysis using Som Classifier Algorithm along with Enhanced Preprocessing Techniques

S. Nagini[1*], T. V. Rajinikanth[2] and B. V. Kiranmayee[1]

[1]Department of CSE, VNR VJIET, Hyderabad - 500090, Telangana, India; nagini_s@vnrvjiet.in, bvk_1973@yahoo.com
[2]Department of CSE, SNIST, Hyderabad - 501301, Telangana, India; rajinitv@gmail.com

Abstract

Objectives: Ages back mankind depends on agriculture yield and it is the only source for food, income and wealth. Even today people of countries like India depend majorly on Agriculture and allied sectors for their livelihood. Most of India's income source is from the agriculture sector. Agriculture yield estimation and analysis are not taking place effectively. **Method/Analysis:** In this regard an algorithm to train the SVM's i.e., the Sequential Minimal Optimization (SMO), classifier algorithm was proposed and results showed that classifier accuracies were improved when compared to other existing techniques. The process involves in replacing all missing values globally. This implementation is globally and then changes nominal attributes to binary form. By default all the attributes are normalized. **Findings:** Classifier coefficients output is purely from normalized data rather than from original data, which is very useful and important. Pair wise coupling is a multi-class classification method. Approach addresses the predicted probabilities that are coupled with the pair-wise coupling method of Hastie and Tibshirani's. The accuracies were very low when 10 fold cross validation is applied. **Novelty/Improvement:** The pre-processing techniques were enhanced to further improve the performance accuracies of SMO algorithm even when cross fold validation is applied on the data sets. Performance based com-parisons were made with the existing techniques.

Keywords: Agriculture Yield, Estimation and Analysis, Multi-Class Problems, Preprocessing Techniques, Sequential Minimal Optimisation

1. Introduction

Even today Agriculture is the backbone of India's economy. To intensify the researchers need and enable them to provide timely the necessary knowledge on crops and their yield in scientific manner on regular basis, John Platt's sequential minimal optimisation algorithm[1] is used. The algorithm SOM is used for training a support vector classifier.

The process involves in replacing all missing values globally this implementation globally and then changes nominal attributes to binary form. By default all the attributes are normalized. Classifier coefficients output is purely from normalized data rather than from original data, which is very useful and important. Pair wise coupling method of Hastie and Tibshirani is a multi-class problem that helps in building logistic models. SOM obtains class membership probability estimates by coupling method for obtaining class membership probability estimates for multi-class classification problems by coupling the probability estimates derived by binary classifier. For obtaining better probability estimates, the method could fit aptly logistic regression models to the SVM's output. SMO uses the Support Vector Machine (SVM)[2,3] for solving Quadratic Programming step that is developed during training in SVM. Lib SVM tool is used for implementing SMO. For misclassified examples, weights can be adjusted to find a linear separator by using

* *Author for correspondence*

Perceptron learning algorithm. Unlike Perceptron, SMO maintains cumulative sum over example weight times the labels. Hence for each label SMO repeatedly adjusts weights.

In[4] discussed about most popularly used data mining techniques like KNN, ANN, k-means clustering and SVM. They stated that Data mining techniques can easily address the most complex agriculture related problems. SVM a binary classifier[5-7] derives two disjoint classes from the data samples. The idea is to have two linearly separable classes for consideration; a hyper plane does this separation. Many hyper planes do this separation, but the one with best margin creation is chosen as classifier. Misclassification will be less when the margin is high.

In[8] stated that the best yield/forecast/estimate can be obtained from Bayesian hierarchical model, as the uncertainty is easily quantified. It combines information from multiple surveys conducted at regular time intervals based on different field measurements, temporal supports and farmers. They found that hierarchical model produces superior forecasts.

In[9] stated that crop yield prediction is a complex task requires various technologies like Statistics, Data Mining and Agriculture. They found that MP5 model tree is the most suitable and effective method in crop yield prediction. It is a combination of classification and regression[10] techniques. It implements top down decision tree method, but in the leaves it has linear regression functions instead of class labels. Decision is made whether to partition the training set or to introduce a regression function as a leaf node.

In[11] stated that SMO is a coherent method for training SVMs on classification tasks. It can handle regression problems. One of the drawbacks of SMO is that its rate of convergence slows down if data is non-sparse, since its caching kernel function output degrades its performance. For regression problems, the modifications improve convergence time by over an order of magnitude.

In[12] stated that effective analysis on yield prediction was done by combining both classification and clustering techniques, a hybrid approach and further it can be extended to various agricultural spatial locations.

In[13] stated that Training a SVM requires a solution for a very large QP optimisation problem. A large QP problem is broken down into a chain of tiny QP problems and is solved analytically consuming less time.

In[2] stated that both quantitative and qualitative methods can be used for the analysis of geographic based agricultural data obtained from certain geographical boundaries. The geographical data is combination of land use profiles and rainfall history which was interpolated to fit into a grid surface. The resultant stochastic yearly rainfall profiles were used to identify areas of high crop yields. It showed that the crop yield is directly proportional to rainfall.

In[14] stated that better crop yield is affected directly by weather parameters. They have used the Artificial Neural Networks method. A suitable crop yield can be predicted by sensing various parameter of soil (like calcium, PH, nitrogen, organic carbon, phosphate, potassium, magnesium, sulphur, manganese, copper, iron) and weather (like rainfall, temperature, humidity, etc.).

2. Materials and Methods

Sequential Minimal Optimization (SMO) is one way to solve the SVM training problem that is more efficient than standard QP solvers. SMO uses heuristics to partition the training problem into smaller problems that can be solved analytically. Whether or not it works well depends largely on the assumptions behind the heuristics (working set selection). Typically, it speeds up training. The Objective function[1] is shown below

$$\min_{\vec{\alpha}} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i \; y_j K\left(\vec{x}_i, \vec{x}_j\right) \alpha_i \alpha_j - \sum_{i=1}^{N} \alpha_i$$
$$0 \leq \alpha_i \leq C, \forall \; i \quad \sum_{i=1}^{N} y_i \; \alpha_i = 0 \tag{1}$$

The SMO searches through the feasible region of the dual problem and maximizes the Objective function. It uses heuristics to choose two for optimization. It is a Hill Climbing Technique. The amount of memory required is linear so it handles very large data sets as it avoids large matrix computation. This SMO algorithm[13] scales somewhere between quadratic and linear for the problems with various training set sizes. This algorithm is the fastest for sparse data sets and linear SVM's i.e., computation time is quite less.

The kernel used is Poly kernel and is commonly used with SVM and SMO. Poly kernel is equivalent to polynomial regression.

$$K(x, y) = (x^{\mathsf{T}} y + c)^{d} \tag{2}$$

where the vectors of the input space are y and x i.e.,

vectors represent features computed from testing and training samples , and $c \geq 0$ is a parameter that influences the lower-order versus higher-order terms of the above polynomial. Kernel is said to be homogeneous when c = 0. Symmetrical Uncertainty[15] Attribute Evaluator filter measures the symmetrical uncertainty with respect to a given class and finally evaluates the value of an attribute. The formula for calculating the value is given below.

$$\text{SymU(Class(C), Attribute(A))} = 2 * (\text{H(Class(C))} - \text{H(Class(C)} \mid \text{Attribute(A))) / H(Class(C)) + H(Attribute(A)))} \quad (3)$$

$$\text{H(Class(C))} = \downarrow \sum_{i=1}^{k} P(Xi) log2\left(P(Xi)\right) \quad (4)$$

H(Class(C)/Attribute(A))=

$$-\sum_{j=1}^{k} P(Yj) \sum_{j=1}^{k} P\left(\frac{Xi}{Yj}\right) Log2\ P\left(\frac{Xi}{Yj}\right) \quad (5)$$

H(Attribute(A))=

$$-\sum_{j=1}^{k} P(Yj) \sum_{j=1}^{k} P\left(\frac{Xi}{Yj}\right) Log2\ P\left(\frac{Xi}{Yj}\right) \quad (6)$$

To find the degree of association among discrete features (Y and X) Symmetrical uncertainty[16] is used. Features values are restricted to be in the range [0, 1]; thereby this compensates the information gain bias[17]. Whether the value of a feature predicts completely the value of another feature is indicated by 1, whereas the independence of X and Y is indicated by 0. Irrespective of this pair of features is treated to be symmetrical.

Ranking an attribute is purely based on each of its individual evaluations; this process is done by Ranker Attribute, their individual. This ranking is used in combining with attribute evaluators like Gain Ratio, Relief F, Entropy etc. For distance measurements k-means cluster uses the Euclidean distance method.

3. Results and Discussion

The selected attributes form the data set after applying the Symmetrical Uncertainty, Attribute Evaluator and Ranking Filter are TTYR, TTPR, TTAT, TTPT, TTAR, TDPR, TDAR, TDPT, TDAT, BLAK, BLAT, TDYR, BLPT, BLPK and BLYK. Two simple clusters were formed using Add cluster filter with simple K-means algorithm on which Euclidean distance is applied. The remaining attributes are removed. Table 1 shows the performance comparisons of classifiers of SVM and SMO Algorithms with respect to various kernels. The three approaches implemented are as follows

- Without Selection of Features (WOSF)
- With selection of Attributes based on Symmetrical Uncertainty Attribute Evaluator with Ranker Algorithm (WSF-SUAE-WRA) and
- With selection of Attributes based on Symmetrical Uncertainty Attribute Evaluator with Ranker Algorithm with ADD Cluster using Simple K-Means Algorithm (WSF-SUAE-WRA-WACSKM). The third approach has proved to be very good in performance when compare to other two approaches. In that Particularly SMO with Poly kernel reached 100% performance after the application of filter ADD Cluster – Simple K-means Algorithm with Euclidean Distance Measure over Feature selection done based on Symmetrical Uncertainty Attribute Evaluator with Ranker Algorithm. The performances were increased after the application of filters in the pre-processing stage over feature selection. The Performance measures used here are Relative Absolute Error, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), F-measure, Time to build Models, Kappa Statistic and percentage of properly classified instances. The graph Classifiers SMO and SVM Comparison using WOSF Approach is represented in Figure 1 here X-axis shows classifiers with kernels and Y-axis shows the performance in terms of numeric values. In this both SMO and SVM performances are almost same except in time taken to build the models. The % of correctly classified instances is zero for the two classifiers SMO and SVM with any kernel. The Kappa statistics value is negative and F-measure values are zeros for all kernels of the two classifiers SMO and SVM. Mean Absolute & Relative Absolute errors are same irrespective of kernels and classifiers. The graph classifiers SMO and SVM comparison using WSF-SUAE-WRA approach is represented by Figure 3 in which X-axis is shown as Classifiers with kernels and Y-axis as performance in terms of Numeric values. In this both SMO and SVM performances are almost same except in time taken to build the models. The % of correctly classified instances is zero for the two classifiers SMO and SVM with any kernel. The Kappa statistics value is negative and F-measure values are zeros for all kernels of the

two classifiers SMO and SVM. Mean Absolute and Relative Absolute errors are same irrespective of kernels and classifiers. The graph Classifiers SMO and SVM Comparison Using WSF-SUAE-WRA-WACSKM Approach is represented in Figure 3 on which X-axis shows Classifiers with kernels and Y-axis shows performance in terms of Numeric values. In this both SMO and SVM performances are almost same except in time taken to build the models. The performance was increased after applying ADD Cluster Filter with Simple K-means Algorithm using Euclidean distance measure before SMO and SVM classifiers were applied. This pre-processing step has drastically enhanced the performances of the classifiers even though 10 fold cross validation is considered. Out of all kernels Poly kernel with SMO proved to be good with highest performance.

Table 1. Performance comparisons of classifiers SVM and SMO with different Kernels apart from feature selection and preprocessed techniques based enhancement

Performance Parameters	Kernel Type	SVM			Kernel Type	SMO		
		WOSF	WSFSUAE-WRA	WSFSUAE-WRAWACSKM		WOSF	WSFSUAE-WRA	WSFSUAE-WRAWACSKM
MAE	Linear	0.0833	0.0833	0.0833	Normalized Poly kernel	0.0833	0.0833	0.3333
RAE		102.2475	102.2475	16.6667		102.2475	102.2475	66.6667
RMSE		0.2887	0.2887	0.2887		0.2099	0.2099	0.5774
F-Measure		0	0	0.917		0	0	0.635
Kappa Statistic		-0.0435	-0.0435	0.8322		-0.0435	-0.0435	0.2993
Time in sec		1.42	2.08	0.05		3.42	0.92	0.92
% of Correctly Classified Instances		0	0	91.6667		0	0	66.6667
MAE	Polynomial	0.0833	0.0833	0.1667	Poly kernel	0.0833	0.0833	0
RAE		102.2475	102.2475	33.3333		102.2475	102.2475	0
RMSE		0.2887	0.2887	0.4082		0.2099	0.2099	0
F-Measure		0	0	0.833		0	0	1
Kappa Statistic		-0.0435	-0.0435	0.669		-0.0435	-0.0435	1
Time in sec		2.5	2.38	0.03		2.38	1.78	0.02
% of Correctly Classified Instances		0	0	83.3333		0	0	100
MAE	RBF	0.0833	0.0833	0.4583	RBF	0.0833	0.0833	0.4583
RAE		102.2475	102.2475	91.6667		102.2475	102.2475	91.6667
RMSE		0.2887	0.2887	0.667		0.2099	0.2099	0.667
F-Measure		0	0	0.381		0	0	0.381
Kappa Statistic		-0.0435	-0.0435	0		-0.0435	-0.0435	0
Time in sec		1.45	2.66	0.03		1.23	1.17	0.03
% of Correctly Classified Instances		0	0	54.1667		0	0	54.1667
MAE	Sigmoid	0.0833	0.0833	0.5	PUK	0.0833	0.0833	0.125
RAE		102.2475	102.2475	100		102.2475	102.2475	25
RMSE		0.2887	0.2887	0.7071		0.2099	0.2099	0.3536
F-Measure		0	0	0.453		0	0	0.872
Kappa Statistic		-0.0435	-0.0435	-0.0511		-0.0435	-0.0435	0.7429
Time in sec		2.56	1.73	0.08		1.44	1.97	0.02
% of Correctly Classified Instances		0	0	50		0	0	87.5

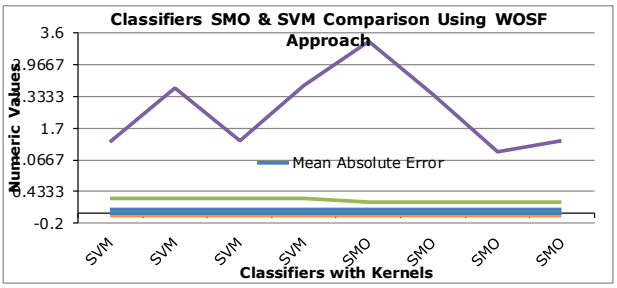

Figure 1. Performance parameters based comparison of WOSF approach.

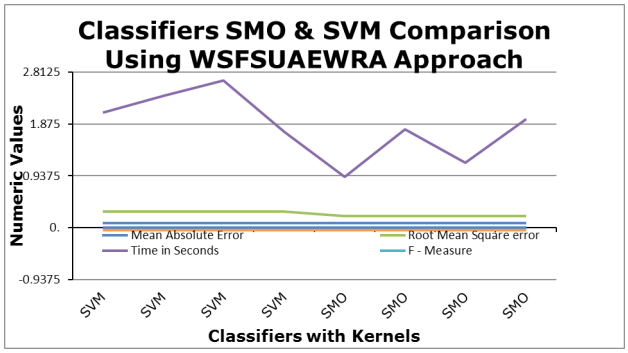

Figure 2. Performance parameters based comparison of WSFSUAEWRA approach.

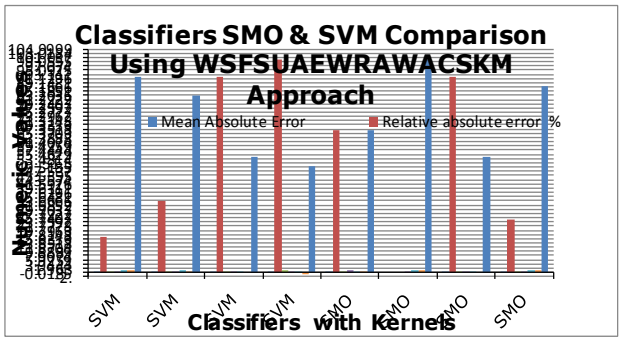

Figure 3. Performance parameters based comparison of WSFSUAEWRAWACSKM approach.

4. Conclusion

The two approaches WOSF and WSF-SUAE-WRA performed almost in the same way. It indicates that even if features were selected by applying Symmetrical Uncertainty Attribute Evaluator with Ranker Algorithm over the data set it does not improved the performance of either SMO or SVM classifiers with whatever kernels considered. Only there is a change in terms of execution time to build models. The results of F-Measure and % of correctly classified instances are coincident in Figure 2 and Figure 3. The Approach WSF-SUAE-WRA-WACSKM proved to be very good in enhancing the performance i.e. % of Correctly classified instances, Kappa Statistic and F-Measure reached highest values whereas the Mean Absolute error and Relative Absolute Error and Root Mean Square error became zero. The application of SMO classifier with Poly kernel proved to be very good after the application of techniques like Feature selection using Symmetrical Uncertainty Attribute Evaluator with Ranker Algorithm and then Add Cluster filter with Simple K-Means using Euclidean Distance.

5. References

1. Platt J. SVM by Sequential Minimal Optimization (SMO). Lecture by David Page. Available from: pages.cs.wisc. edu/ ~dpage/ cs760/ MLlectureSMO.ppt
2. Vagh Y. An investigation into the effect of stochastic annual rainfall on crop yields in South Western Australia. International Journal of Information and Education Technology. 2012 Jun; 2(3).
3. Cao LJ, et al. Parallel sequential minimal optimization for the training of support vector machines. IEEE Transactions on Neural Networks. 2006; 17:1039-49.
4. Mucherino A, Petraq P, Pardalos PM. A survey of data mining techniques applied to agriculture. Oper Res Int J. 2009; 9:121–40. DOI:10.1007/s12351-009-0054-6
5. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov. 1998; 2(2):955–74.
6. Cortes C, Vapnik V. Support vector networks. Mach Learning. 1995; 20:273–97.
7. Vapnik VN. Statistical Learning Theory. New York: Wiley; 1998.
8. Wang JC, Holan SH, Nandram B, Barboza W, Toto C, Anderson E. A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. Journal of Agricultural, Biological and Environmental Statistics. 2012; 17(1):84–106. DOI:10.1007/s13253-011-0067-5.
9. Marinkovic B, et al. Data mining approach for predictive modeling of agricultural yield data.
10. Quinlan JR. Learning with continuous classes. Proceedings of 5th Australian Joint Conference on Artificial Intelligence; 1992. p. 343-8.
11. Flake GW, Lawrence S. Efficient SVM regression training with SMO*. Machine Learning. Kluwer Academic Publishers; 2002; 46:271–90.
12. Rao ChM, Rao AA. Crop yield analysis of the irrigated areas of all spatial locations in Guntur District of AP. 2014 Jun; 4(6):1-7.

13. Platt JC. Sequential minimal optimization: A fast algorithm for training support vector machines.
14. Dahikar SS, Rode Sandeep V. Agricultural crop yield prediction using artificial neural network approach. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering. 2014 Jan; 2(1).
15. Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research. 2004; 5:1205–24.
16. Hall MA, Holmes G. Benchmarking attributes selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering. 2003 May/Jun; 15(3).
17. Liu H, Hussain F, Tan CL, Dash M. Discretization: An enabling technique. Data Mining and Knowledge Discovery. 2002; 6(4):393–423.