American Journal of Applied Sciences, 2012, 9 (12), 1934-1939
ISSN: 1546-9239

doi:10.3844/ajassp.2012.1934.1939 Published Online 9 (12) 2012 (http://www.thescipub.com/ajas.toc)

A FUZZY FAST CLASSIFICATION FOR SHARE MARKET DATABASE WITH LOWER AND UPPER BOUNDS

[1]Srinivasan Vaiyapuri, [2]Rajenderan Govind and [1]Vandar Kuzhali Jaganathan

[1]Department of Computer Applications, Velalar College of Engineering and Technology,
[2]Department of Science and Humanities, Kongu Engineering College,
Erode, Tamilnadu-638012, India

Received 2012-07-09, Revised 2012-08-24; Accepted 2012-11-08

ABSTRACT

In recent years, many researchers focused on the research topic of constructing fuzzy classification system. This study introduces a Fuzzy Fast Classification (FFC) approach for large data sets. It has three phases, in the first phase the large data base is reduced with the entropy by removing the number of attribute. In the second phase an approximate classification is obtained by the mean separation of the data by the total weight, upper and lower approximation line is drawn such that 20% of the record lies near the mean line. In the third phase the classification is refined by using fuzzy logic approach for the 20% of the record since they may fall in any one of the category which need to be carefully examined with the degree of fuzzy value. Experimental results for share market database demonstrate that our approach has good classification accuracy while the training is significantly faster than other SVM classifiers. The proposed classifier has distinctive advantages on dealing with huge data sets.

Keywords: Classification, Entropy, Information Gain, SVM, Fuzzy SVM, Fuzzy Fast Classification

1. INTRODUCTION

The theory of Support Vector Machines (SVM) is a classification technique that has drawn much attention in the recent years (Xuegong, 2000; Jianhua and Gong, 2011; Dong *et al*., 2005; Holte, 1993). SVM has shown to provide higher performance than traditional learning machines and has been introduced as a powerful tool for solving classification problems. There are more and more applications using the SVM techniques (Yu *et al*., 2003). However, in many applications, some input points may not be exactly assigned to one of these two classes exactly. Some are more important to be fully assigned to one class so that SVM can separate these points more correctly. Some of the applications using the support vector techniques with different problems (Chen and Wang, 2003). In the recent years we need efficient algorithms that reduce the time for classification. In this study, we apply a fuzzy approach for fast classification to input point that lies near the mean line which separates the given database into two classes. Now we reformulate the SVM into Fuzzy Fast Classification such that the input points that lie near the mean line are more difficult to classify with the learning of the decision surface. The proposed method enhances the SVM by using lower and upper bound line to reduce the time in data points to find the classes. We execute our proposed approach with the real time database for share market, various research on share market classification with different techniques and problem are seen in (Vaisla and Bhatt, 2010; Walczak, 2001; Wang *et al*., 2008). The rest of this study is organized as follows. We present how the attribute selection is more important to reduce the database. Next we give brief review of the theory of SVM and finally implement our new approach fuzzy fast classification, with experiments result and concluding remarks.

2. MATERIALS AND METHODS

2.1. Entropy

Entropy is a measure of variability in a random variable. It will measure how the particular attribute divides the training examples into the number of result

Corresponding Author: Srinivasan Vaiyapuri, Department of Computer Applications, Velalar College of Engineering and Technology, Erode, Tamilnadu-638012, India

classes (Jing *et al*., 2007; Halperin and Kar, 2005). Table 1 show the sample database for the share market data. In our problem we need to select the attribute which gives more information for classification so as to make accurate classification. For defining gain entropy is obtained from information theory. Entropy is used to calculate the amount of useful information in an attribute. This is calculated as Equation 1:

$$\text{Entropy}(S) = -\Sigma P(x_i) \log_b P(x_i) \quad (1)$$

Where:
S = Collection of Samples
x_i = Set of outcomes
$P(x_i)$ = Proportion of S to the class x_i

Entropy was used by J. Ross Quinlan who has used Entropy in ID3 algorithm. This algorithm is based on the Concept Learning System (CLS).

2.2. Information Gain (IG)

The information gain is based on the decrease in entropy after a dataset is split on an attribute. First the attribute that creates the most homogeneous branches are identified Equation 2:

$$IG(Y/X) = H(Y) - H(Y/X) \quad (2)$$

The entropy calculated for the database of share market database are ordered as public share holding with 0.97, high-average 0.92, 52 weeks high 0.88, Average of 52 weeks 0.86 and so on..

Table 1. Sample database of partial record is shown from the Indian share market database

A1	A2	A3		…	A15	A16
1	Adani enterprises	1.00	1.00		1.00	2
2	Adityabirla ro1vo	0.03	0.06		0.04	1
3	Andhra bank	0.07	0.06		0.07	1
	…			…		
48	Union bank of India	0.12	0.13		0.13	2
49	United phos phorous	0.03	0.04		0.03	2
50	Yes bank	0.03	0.04		0.04	2

A1-S.No, A2-Share name, A3-Q1-Profit from April to June, A-Yearly profit, A5-52_weaks high, A6-Average of q1 to q4, A7-Average profit of the year, A8-Average of 52 weeks, A9-Q2-profit from July to September, A10-Q3-Profit from October to December, A11-Q4-Profit from January to March, A12-Public share holding, A13-52 weeks low, A14-High-Average, A15-Average total, A16-ClassA14-High-Average, A15-Average total, A16-Class

2.3. Support Vector Machine (SVM)

Support vector machines are learning machines that can perform binary classification with regression estimation tasks. SVMs are also recognized as efficient tools for data mining and are popular because of two important factors. First, unlike the other classification techniques, SVMs minimize the expected error rather than minimizing the classification error. Second, SVMs employ the duality theory of mathematical programming to get a dual problem that admits efficient computational methods of SVM (Leng and Wang, 2008). SVM incorporate structured risk minimization into the classification. By structured risk minimization, we mean minimizing upper and lower bound on the generalization error. Consider a simple case when two data sets, A and B, are linearly separable. Traditionally, we attempt to discriminate the points in A and B by constructing a separating line $X' W = \gamma$, so that the open half space $\{X \mid X \in R^n, X' W > \gamma\}$ contains mostly the points of A and the other open half space $\{X \mid X \in R^n, X' W < \gamma\}$ contains mostly the points of B. In other words, we wish to determine W and γ, such that the following two inequalities are satisfied:

$$AW > e\gamma \text{ and } BW < e\gamma$$

where, A denotes the matrix corresponding to all X in class A; B denotes a matrix for data in B; and e is the vector of all 1s. We can write the normalized version of the above pair of inequalities as:

$$AW > e\gamma + e \text{ and } BW < e\gamma - e$$

This means that for every X in the class A, we have $X' W - \gamma < +1$ and for every X in class B, we have $X' W - \gamma < -1$. **Figure 2** shows the division of members into two classes for the Indian stock market.

2.4. Fuzzy Logic

Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth, truth values between completely true and completely false. It was introduced by Dr. Lotfi Zadeh in the 1960's as a means to model the uncertainty of natural language.

2.5. Fuzzy Subsets

In classical set theory, a subset U of a set S can be defined as a mapping from the elements of S to the elements of the set {0, 1}, U: S --> {0, 1} This mapping may be represented as a set of ordered pairs, with exactly

one ordered pair present for each element of S. The first element of the ordered pair is an element of the set S and the second element is an element of the set {0, 1}. The value zero is used to represent non-membership and the value one is used to represent membership. The truth or falsity of the statement X is in U is determined by finding the ordered pair whose first element is X. The statement is true if the second element of the ordered pair is 1 and the statement is false if it is 0.

Similarly, a fuzzy subset F of a set S can be defined as a set of ordered pairs, each with the first element from S and the second element from the interval [0,1], with exactly one ordered pair present for each element of S. This defines a mapping between elements of the set S and values in the interval [0, 1]. The value zero is used to represent complete non-membership, the value one is used to represent complete membership and values in between are used to represent intermediate Degrees of Membership.

In our problem the set S is the set of attributes. Let's define a fuzzy subset High, which will give us the degree x is high. To each plant in the universe of discourse, we have to assign a degree of membership in the fuzzy subset High. The membership function is formed based on the attribute value X and in the same way membership function is framed for all the four attributes for our iris database and for the share market database, **Fig. 1** sample is given below for one attribute, similarly it is done for all the attributes in the database:

S.L(x) = {0, if S.L(x) <= 0.73,
(S.L(x)-0.74)/0.07, if 0.74 < S.L(x) < 0.80,
1, if S.L(x) >= 0.80}

2.6. Logic Operations

The standard definitions in fuzzy logic are:

truth (not x) = 1.0 - truth (x)
truth (x and y) = minimum (truth(x), truth(y))
truth (x or y) = maximum (truth(x), truth(y))

Assume that the variables A1, A2, A3 and R all take on values in the interval [0,10] and that the following membership functions and rules are defined:

Low (t) = 1 - (t / 10)
High (t) = t / 10
Rule 1: if A1 is and A2 is L and A3 is L then R is L
Rule 2: if A1 is L and A2 is L and A3 is H then R is L
Rule 3: if A1 is L and A2 is H and A3 is L then R is L
Rule 4: if A1 is L and A2 is H and A3 is H than R is H
Rule 5: if A1 is H and A2 is L and A3 is L then R is L
Rule 6: if A1 is H and A2 is L and A3 is H then R is H
Rule 7: if A1 is H and A2 is H and A3 is L then R is H
Rule 8: if A1 is H and A2 is H and A3 is H then R is H
L – Low, H – High

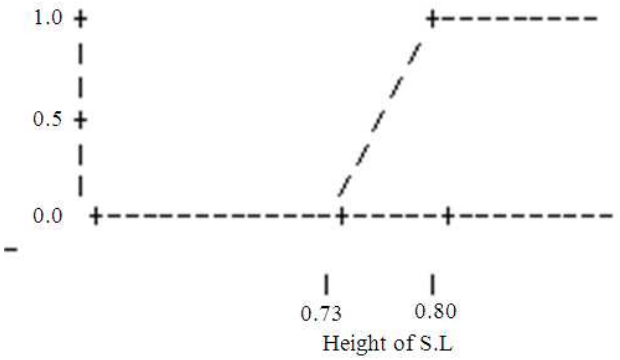

Fig. 1. Shows the Fuzzy values for sepal length

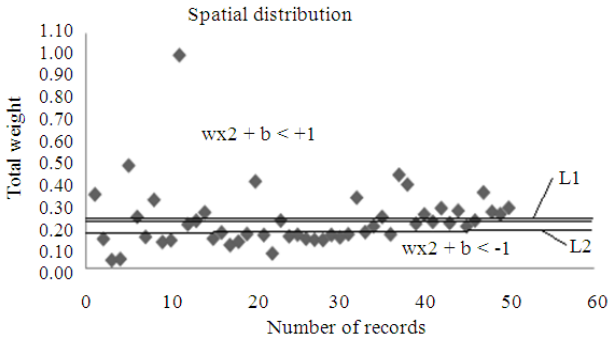

Fig. 2. Shows the Linear line mean separation with upper and lower bound for the share market database

2.7. Implementation of FFC

In SVM methodology, there will be just one optimal line: the line lying half way in between the maximal margin of classifying into high and low. To find this exact line that separates the two part of the plane is hard to find and it consumes more time (Tsang *et al*., 2003). **Figure 2** shows that FFC classifies the members based on the mean line by avoiding the continuous training to find the optimal line. Taking the mean as the optimal line, we further go for fuzzy approach with small displacement to this optimal line such that the records lie between this lower an upper bound is 20% of the total records. The member that lies between the upper and lower bound alone is taken special care in classifying the members by analyzing the each attribute based on the entropy hierarchy with a fuzzy approach. With this implementation of fuzzy fast classification with the mean line reduces the time that spend on finding the optimal line in SVM makes our FFC with high speed and more accuracy.

The linear classifier is the mean line L (y = w * x + b) with the maximum width (distance between linear line L_1 and L_2). So the numbers of records that comes between these two line are less than or equal to 20%. The records that lie between these two lines are classified by

checking each attribute order based on the entropy hierarchy because we have considered the case of linearly separable classes, each such linear line (w, b) is a classifier that approximately separates all patterns from the training set:

$$\text{Class}(x_i) = \begin{cases} +1 \text{ if } w' x_i + b > 0 \\ -1 \text{ if } w' x_i + b < 0 \end{cases}$$

For all points from the linear line L (w * x +b = 0) the distance between origin and the linear line L is |b|/||w||. We consider the patterns from the class -1 that satisfy the equality w * x + b = -1 and determine the linear line L_1; the distance between origin and the linear line L_1 is equal to |-1-b|/||w||. Similarly, the patterns from the class +1 satisfy the equality w * x + b = +1 and determine the linear line L_2; the distance between origin and the linear line L_2 is equal to |+1-b|/||w||. The linear lines L, L_1 and L_2 are parallel and no training patterns are located between linear lines L_1 and L_2. Based on the above considerations, the distance between linear lines (margin) L_1 and L_2 is 2/||w||. The improved FFC is done by fixing the lower and upper bound line. That is we divide the members into two classes by the mean line and argue that errors are likely to occur near the mean line. The division of this membership function is carried out by the mean line to classify low and high but we use fuzzy logic approach for the records that lie near the mean line which need to be carefully classified with our new approach FFC. In our problem we have two classifications of low and high. The two categories of low and high is obtained easily by using the mean line with two categories, But we argue that the error caused by the classification is more near the mean line, So upper and lower lines are framed such that 20% of record near the mean line is obtained as a error records that need to be carefully classified with a fuzzy logic approach.

2.8. Proposed Algorithm

1. Let v be the number of record
2. TV =Calculate Total value for each record
3. M = Find the mean of TV
4. 20% of the records near the mean line.
5. L1 is fixed such that low category has > than or equal to 40% of the records from total records.
6. L2 is fixed such that high category has > than or equal to 40% of the record from total records.
7. Vectors above L1 is classified as high
8. Vectors below L2 is classified as low
9. For (I = 0; i <n; i ++)
10. If ((v(i).TV) between L2 && L1))
11. Go for fuzzy logic approach to classify
12. End

With the help of mean separation, maximum of the records are classified as low and high that is 80% of the records are classified, except for the complex area near the mean line that is 20% of the record need to be classified examining each hierarchy attribute by the entropy Equation 1 and the information gain Equation 2. With this new approach fuzzy fast classification is obtained by analysing the degree of fuzziness for each attribute and finally classified with a fuzzy approach with more fast and accurate classification.

3. RESULTS

We evaluate the performance of the proposed FFC by applying it to the Iris database (Mertz and Murphy, 1998). These data sets are well-Known benchmark data for evaluating the performance of classifiers. The Iris database created by fisher includes three classes with 150 instances and 50 records for each category. As we deal with binary classification we consider any two categories at a time and also show our result with the different combination. In our example the iris database is classified with two classes using the fuzzy logic approach, The separation of the mean line is done with average of the total weight, the upper and lower lines are framed in connection with the mean line based on the condition that 20% of the records that lies between the lower and the upper bound, theses records are considered as the miss classified records and need to be classified based on our new fuzzy approach.

We apply our algorithm with the real time database of Indian share market. The mean line occurs at 0.22. The L1 is set at 0.22 and the L2 is set at 0.16 which is optimal to get 20% of the error records. The records that lie between these two lines may be misclassified. In the share database 7 error records lies between L1 and L2 which are shown in **Table 2**. These seven records are correctly classified with our mean line separation itself so we need not go for fuzzy approach. If any error records arises than we go for fuzzy logic approach to correct the errors as we did in the iris database. With this new approach we can see that our approach makes the classification more fast with accurate classification.

The experiment result shows that our proposed method FFC has significant advantage over the SVM classifier, by finding the optimal line for the SVM which consumes more time to undergo many training and then to find the optimal line which reduces the speed of the SVM, taking this into the consideration and to find the optimal line for FFC makes it simple by mean line separation with the fuzzy logic approach to solve the error records.

Table 2. Shows 7 records are misclassified and classified correctly by the hierarchy of the attributes

Share name	Public share holding	Diff. high-Avg	52 weeks high	AVG of 52 weeks	Q4	52 weeks low	Total weight	Class	Name class
Cummins India	0.36	0.14	0.08	0.07	0.30	0.06	0.17	1	1
Lic Hsg Fin	0.53	0.05	0.03	0.03	0.60	0.03	0.17	1	1
Titan Inds	0.69	0.05	0.03	0.03	0.02	0.03	0.19	1	1
Lupin	0.54	0.07	0.07	0.06	0.04	0.06	0.20	1	1
Canara bank	0.30	0.19	0.09	0.07	0.17	0.06	0.20	1	1
Corp tata	0.59	0.10	0.04	0.03	0.11	0.02	0.21	1	1
Chemicals	0.65	0.07	0.05	0.05	0.04	0.05	0.21	1	1

Table 3. Show the performance and time evaluation

Algorithm	Performance	Time
FFC (iris)	0.98	0.14
SVM (iris)	0.96	0.17
FFC (share)	0.99	0.10
SVM (Share)	0.98	0.13

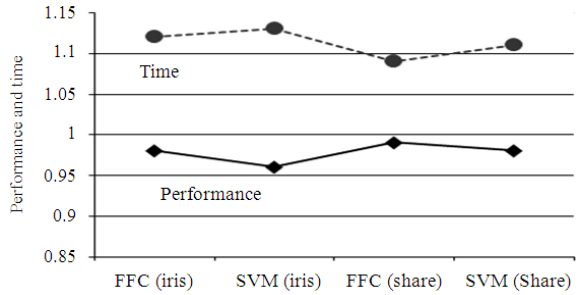

Fig. 3. Shows the performance of the two algorithms with performance and time

We are able to classify the maximum record into two classes with mean line separation. The misclassified records are very less which can be classified by checking by the hierarchy of attribute selection based on the entropy and the information gain with a fuzzy logic approach.

Thus by using the fuzzy approach the proposed method FFC removes the drawback of the SVM algorithm and shows good performs for our proposed method with fast and high accuracy classification. These experiments were carried out in the MATLAB and the results are shown in the **Table 3 and Fig. 3**.

4. CONCLUSION

In this study we have proposed a new approach FFC for fast classification with more accuracy. The efficiency of the SVM algorithm is low due do the time consuming process to finding the exact optimal line to separate the binary members. FFC solves this problem by using the mean line separation for a large database which reduces the time to find the optimal line. With this mean line we obtain 20% of the records that lie near the mean line need to be carefully assigned the class as the error or misclassification normally occurs near the mean line. The upper and lower lines are framed such that the 20% of the records that lies near the mean. The record that comes between theses two lines are 19 records for iris database and 7 records for the share database. Out of 19 records 15 records are correctly classified by mean line and rest of the 4 records is to be carefully examined based on the hierarchy of the attributed with entropy and finally classified with our new approach using the fuzzy logic operation with the degree of fuzziness. The experimental results show that using FFC approach has improved the performance of the algorithm with the training time significantly reduced with the proposed approach, which can be scalable to huge data sets to obtaining fast classification with high classification accuracy.

5. REFERENCES

Chen, Y. and J.Z. Wang, 2003. Support vector learning for fuzzy rule-based classification systems. IEEE Trans. Fuzzy Syst., 11: 716-728. DOI: 10.1109/TFUZZ.2003.819843

Dong, J.X., A. Krzyzak and C.Y. Suen, 2005. Fast SVM training algorithm with decomposition on very large data sets. J. IEEE Trans. Patt. Anal. Mach. Intell., 27: 603-618. DOI: 10.1109/TPAMI.2005.77

Halperin, E. and R.M. Kar, 2005. The minimum-entropy set cover problem. Theory Comput. Sci., 348: 240-250. DOI: 10.1016/j.tcs.2005.09.015

Holte, R.C., 1993. Very simple classification rules perform well on most commonly used datasets. J. Mach. Learn., 11: 63-91. DOI: 10.1023/A:1022631118932

Jianhua, X.Z. and X. Gong, 2011. The new development in support vector machine algorithm theory and its application. J. Control Decision-Mak., 19: 481-484.

Jing, L., M.K. Ng and J.Z. Huan, 2007. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Trans Knowl. Data Eng., 19: 1026-1041. DOI: 10.1109/TKDE.2007.1048

Leng, X.M. and Y.D. Wang, 2008. Gender classification based on fuzzy SVM. Proceedings of the 2008 International Conference on Machine Learning and Cybernetics, Jul. 12-15, IEEE Xplore Press, Kunming, pp: 1260-1264. DOI: 10.1109/ICMLC.2008.4620598

Mertz, C.J. and P.M. Murphy, 1998. UCI Repository of machine learning databases. CiteULike.

Tsang, C.C.C., D.S. Yeung and P.P.K. Chan, 2003. Fuzzy support vector machines for solving two-class problems. Proceedings of the 2003 International Conference on Machine Learning and Cybernetics, Nov. 2-5, IEEE Xplore Press, pp: 1080-1083. DOI: 10.1109/ICMLC.2003.1259643

Vaisla, K.S. and A.K. Bhatt, 2010. An analysis of the performance of artificial neural network technique for stock market forecasting. Int. J. Comput. Sci. Eng., 2: 2104-2109.

Walczak, S., 2001. An empirical analysis of data requirements for financial forecasting with neural networks. J. Manage. Inform. Syst., 17: 203-222.

Wang, L.M., Y.C. Liang, X.M. Han, X.H. Shi and M. Li, 2008. Multi-winners SOMs and their applications to stock analysis. J. Comput. Res. Dev.

Xuegong, Z., 2000. Introduction to statistical learning theory and support vector machines. Journal Acta Autom Sinica. 26: 32-42.

Yu, H., J. Yang and J. Han, 2003. Classifying large data sets using SVMs with hierarchical clusters. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 24-27, ACM Press, Washington, DC, USA., pp: 306-615. DOI: 10.1145/956750.956786