Support vector machines based on K-means clustering for real-time business intelligence systems

Jiaqi Wang*

Faculty of Information Technology,
University of Technology, Sydney,
P.O. Box 123, Broadway, NSW 2007, Australia
E-mail: jqwang@it.uts.edu.au
*Corresponding author

Xindong Wu

Department of Computer,
University of Vermont,
33 Colchester Avenue/351 Votey Building,
Burlington, Vermont 05405, USA
E-mail: xwu@cs.uvm.edu

Chengqi Zhang

Faculty of Information Technology,
University of Technology, Sydney,
P.O. Box 123, Broadway, NSW 2007, Australia
E-mail: chengqi@it.uts.edu.au

Abstract: Support vector machines (SVM) have been applied to build classifiers, which can help users make well-informed business decisions. Despite their high generalisation accuracy, the response time of SVM classifiers is still a concern when applied into real-time business intelligence systems, such as stock market surveillance and network intrusion detection. This paper speeds up the response of SVM classifiers by reducing the number of support vectors. This is done by the K-means SVM (KMSVM) algorithm proposed in this paper. The KMSVM algorithm combines the K-means clustering technique with SVM and requires one more input parameter to be determined: the number of clusters. The criterion and strategy to determine the input parameters in the KMSVM algorithm are given in this paper. Experiments compare the KMSVM algorithm with SVM on real-world databases, and the results show that the KMSVM algorithm can speed up the response time of classifiers by both reducing support vectors and maintaining a similar testing accuracy to SVM.

Keywords: business intelligence; data mining; testing accuracy; response time; classifier; SVM; K-means; KMSVM.

Reference to this paper should be made as follows: Wang, J., Wu, X. and Zhang, C. (2005) 'Support vector machines based on K-means clustering for real-time business intelligence systems', *Int. J. Business Intelligence and Data Mining*, Vol. 1, No. 1, pp.54–64.

Biographical notes: Jiaqi Wang is now a PhD candidate of the Faculty of Information Technology at the University of Technology, Sydney, Australia. His research interests include data mining and high-frequency finance.

Xindong Wu is a Professor and the Chair of the Department of Computer Science at the University of Vermont, USA. He holds a PhD in artificial intelligence from the University of Edinburgh, Britain. His research interests include data mining, knowledge-based systems, and web information exploration. He has published extensively in these areas in various journals and conferences, including IEEE TKDE, TPAMI, ACM TOIS, IJCAI, AAAI, ICML, KDD, ICDM, and WWW. He is the winner of the 2004 ACM SIGKDD Service Award.

Chengqi Zhang is a Research Professor of the Faculty of Information Technology at the University of Technology, Sydney, Australia. He received a PhD degree from the University of Queensland, Brisbane in Computer Science and a Doctor of Science (higher doctorate) degree from Deakin University, Australia. His areas of research are data mining and multi-agent systems. He has published more than 200 refereed papers, edited nine books, and published three monographs.

1 Introduction

The objective of business intelligence (BI) is to make well-informed business decisions by building both succinct and accurate models based on massive amounts of practical data. There are many kinds of models built for different practical problems, such as classifiers and regressors. This paper mainly discuses the related issues about the design of classifiers applied into BI systems. For example, a model can be represented by a classifier when the telecommunication companies predict whether or not their clients probably pay later than required.

Generalisation accuracy and response time are two important criteria for evaluating classifiers when applied into real-time BI systems. Classifiers are required not only to describe training data but also to be able to predict unseen data. In the example about telecommunication companies, the classifiers are expected not only to describe behaviours of current customers, but also, more importantly, to predict behaviours of new customers. Generalisation accuracy can usually be estimated only by some methods because the distribution of data is often unknown and the true accuracy (generalisation accuracy) cannot be calculated. Estimated accuracy is called testing accuracy in this paper. In addition, the response speed of classifiers is expected to be high when applied into real-time BI systems, e.g., in stock market surveillance and network intrusion detection. Even users sometimes may sacrifice a little testing accuracy of classifiers in order to speed up the response of classifiers in real-time BI systems.

Many machine learning algorithms have been developed to improve testing accuracy of classifiers. Among them, one of the most effective algorithms is support vector machines (SVM) proposed by Boser et al. (1992) and Vapnik (1998). Due to its solid mathematical foundation and high testing accuracy, SVM has been widely applied to build classifiers in many applications, e.g., images (Osuna et al., 1997), speech (Ganapathiraju, 2004), text (Joachims, 1998), and bioinformatics (Furey et al., 2000).

Besides, some industrial leaders of data mining have embedded or are embedding SVM into their products, e.g., 'Oracle Data Mining Release 10g'.

Although SVM can build classifiers with high testing accuracy, the response time of SVM classifiers still needs to improve when applied into real-time BI systems. Two elements affecting the response time of SVM classifiers are the number of input variables and that of the support vectors. While Viaene et al. (2001) improve response time by selecting parts of input variables, this paper tries to improve the response time of SVM classifiers by reducing support vectors.

Based on the above motivation, this paper proposes a new algorithm called K-means SVM (KMSVM). The KMSVM algorithm reduces support vectors by combining the K-means clustering technique and SVM. Since the K-means clustering technique can almost preserve the underlying structure and distribution of the original data, the testing accuracy of KMSVM classifiers can be under control to some degree even though reducing support vectors could incur a degradation of testing accuracy.

In the KMSVM algorithm, the number of clusters is added into the training process as the input parameter except the kernel parameters and the penalty factor in SVM. In unsupervised learning, e.g., clustering, usually the number of clusters is subjectively determined by users with domain knowledge. However, when the K-means clustering technique is combined with SVM to solve the problems in supervised learning, e.g., classification, some objective criteria independent of applications can be adopted to determine these input parameters.

In supervised learning, determining the input parameters is called model selection. Some methods about model selection have been proposed, e.g., the hold-out procedure, cross-validation (Langford, 2000), and leave-one-out (Vapnik, 1998). This paper adopts the hold-out procedure to determine the input parameters for its good statistical properties and low training costs. In addition, searching strategies are needed and among them grid search is the most popular one (Hsu et al., 2003). The computational cost of grid search is high when it is used to determine more than two input parameters. Based on grid search, this paper gives a more practical heuristic strategy to determine the number of clusters.

The experiment on the Adult-7 data shows that in the input parameter space constructed only by the kernel parameter γ and the penalty factor C, it is very difficult for SVM to find the combination of input parameters greatly reducing support vectors. The KMSVM algorithm searches the combination of input parameters in the higher-dimensional space constructed by the kernel parameter γ, the penalty factor C, and the number of clusters. Our experiments show that the KMSVM algorithm can find a good combination of input parameters greatly reducing support vectors in this higher-dimensional space and maintaining a similar testing accuracy (the similar testing accuracy in this paper means a 2–3% discrepancy of classification accuracy on the testing data, see Tables 4 and 5) to SVM. For example, on the Adult-7 database, SVM builds a classifier with about 6000 support vectors while the KMSVM algorithm reduces it to about only 100 support vectors with a similar testing accuracy (84.9% vs. 82.1%). Moreover, the response of the KMSVM classifier is about 25 times faster that of the SVM classifier in this experiment.

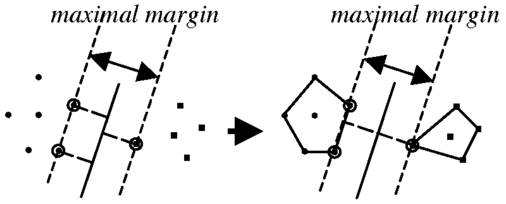
The remainder of this paper is organised as follows. Section 2 introduces SVM and proposes the KMSVM algorithm. Section 3 discuses model selection in the KMSVM algorithm and a heuristic strategy to determine the number of clusters. The experiments on some real-world databases verify the effectiveness of the KMSVM algorithm and the heuristic searching strategy in Section 4. Section 5 draws a conclusion about the KMSVM algorithm.

2 SVM and KMSVM

The theory and algorithm about SVM are originally established by Vapnik (1998) and have been applied to solve many practical problems since 1990s. SVM benefits from two good ideas: maximising the margin and the kernel trick. These good ideas can guarantee high testing accuracy of classifiers and overcome the problem about curse of dimensionality. In classification, SVM solves the quadratic optimisation problem in equation (1), and this optimisation problem is geometrically described in Figure 1.

$$\begin{aligned} &\min \|w\|^2 + C \sum_{i=1\ldots l} \xi_i \\ &s.t. \quad (\{w, x_i\} + b) y_i > 1 - \xi_i,\ \xi_i \geq 0, \quad i = 1 \ldots l \end{aligned} \tag{1}$$

Figure 1 SVM maximises the margin between two linearly separable sample sets. The maximal margin is equal to the shortest distance between two disjoint convex hulls spanned by these two sample sets

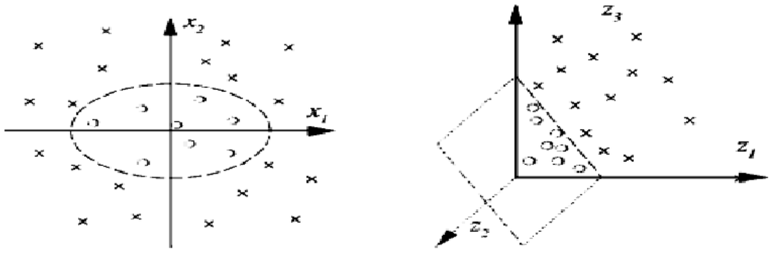


Source: Wang (2002).

In addition, kernel functions are used in SVM to solve non-linear classification problems. Now the popular kernel functions include the polynomial function, the Gaussian radius basis function (RBF), and the sigmoid function. An example of solving the non-linear classification problem is described in Figure 2. SVM classifiers are represented as the formula in equation (2),

$$f(x) = \operatorname{sgn}(\textstyle\sum_{i=1\ldots n} \alpha_i K(x, x_i) + b) \tag{2}$$

Figure 2 The samples are mapped from a 2-dimensional space to a 3-dimensional space. A non-linear classification is converted into a linear classification by using feature mapping and kernel functions



Source: Müller et al. (2001).

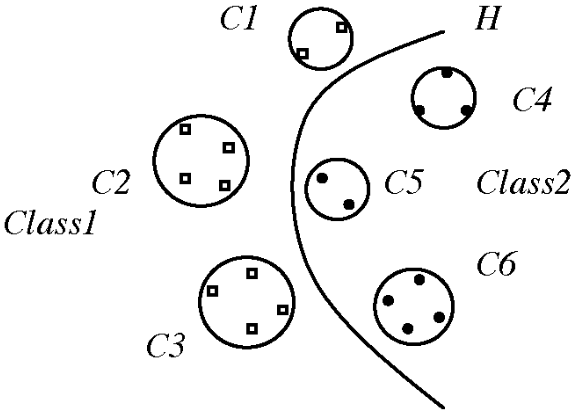
where n is the number of support vectors, x_i is a support vector, α_i is the coefficient of the support vector, b is the bias, $K(\bullet,\bullet)$ is the kernel function, and sgn(•) is the sign function. This paper adopts the Gaussian RBF kernel in equation (3).

$$K(x_1, x_2) = \exp(-\gamma \times \|x_1 - x_2\|^2) \tag{3}$$

The response time of SVM classifiers needs to improve when applied into real-time BI systems. Two main elements affecting the response time of SVM classifiers are the number of input variables and that of support vectors. While Viaene et al. (2001) speed up the response of SVM classifiers by selecting parts of input variables, the response time may still not be acceptable because of too many support vectors, e.g., several thousands of support vectors in Experiment 1. This paper tries the K-means clustering technique to reduce support vectors.

K-means is a classical clustering algorithm in the field of machine learning and pattern recognition (Duda and Hart, 1972). It can almost preserve the underlying structure and distribution of the original data. An example of the K-means clustering technique is shown in Figure 3. There are twenty positive and negative samples in this example, and they are compressed by the K-means clustering technique to only six cluster centres *C1*, *C2, …, C6*. The statistical distribution and structure of the original data are almost preserved when represented by these six cluster centres. This implies that there may be a lot of redundant information in the original data set, and it is possible to build classifiers with acceptable testing accuracy based on those cluster centres compressed by the K-means clustering technique.

Figure 3 K-means clustering is run on the 2-dimensional data set and six clusters are formed, *C1*, *C2*, …, *C6*. *H* is the classifier separating two-class samples into *Class1* and *Class2*



This paper combines the K-means clustering technique with SVM to build classifiers, and the proposed algorithm is called KMSVM. It is possible for the KMSVM algorithm to build classifiers with many fewer support vectors and higher response speed than SVM classifiers. Moreover, testing accuracy of KMSVM classifiers can be guaranteed to some extent. The details of the KMSVM algorithm are described in Table 1.

Table 1 Steps of the KMSVM algorithm

Step 1: three input parameters are selected: the kernel parameter γ, the penalty factor C, and the compression rate *CR*
Step 2: the K-means clustering algorithm is run on the original data and all cluster centres are regarded as the compressed data for building classifiers
Step 3: SVM classifiers are built on the compressed data
Step 4: three input parameters are adjusted by the heuristic searching strategy proposed in this paper according to a tradeoff between the testing accuracy and the response time
Step 5: return to Step 1 to test the new combination of input parameters and stop if the combination is acceptable according to testing accuracy and response time
Step 6: KMSVM classifiers are represented as the formula in equation (2)

3 Model selection

Model selection in the KMSVM algorithm is to decide three input parameters: the RBF kernel parameter γ, the penalty factor C, and the compression rate CR in equation (4) (searching CR is equivalent to doing the number of clusters when the number of the original data is fixed). This section discuses model selection from two perspectives: the generalisation accuracy and response time of classifiers applied into real-time BI systems. Tradeoff of generalisation accuracy and response time determines the values of input parameters.

$$\text{CR} = \textit{No. of original data/No. of clusters} \tag{4}$$

In model selection, generalisation accuracy is usually estimated by some procedures, e.g., hold-out, *k*-fold cross-validation, and the leave-one-out procedure, since the distribution of the original data is often unknown and the actual error (generalisation error) cannot be calculated. The hold-out procedure divides the data into two parts: the training set on which classifiers are trained, and the testing set on which the testing accuracy of classifiers is measured (Langford, 2000). The *k*-fold cross-validation procedure divides the data into *k* equally sized folds. It then produces a classifier by training on $k - 1$ folds and testing on the remaining fold. This is repeated for each fold, and the observed errors are averaged to form the *k*-fold estimate (Langford, 2000). This procedure is also called leave-one-out when *k* is equal to the number of trained data.

This paper recommends and adopts the hold-out procedure to determine the input parameters in the KMSVM algorithm (and SVM) for two reasons. Regardless of the learning algorithms, Hoffding bounds can guarantee that with high probability discrepancy between estimated error (testing error) and true error (generalisation error) will be small in the hold-out procedure. Moreover, it is very time consuming for the *k*-fold cross-validation or the leave-one-out procedure to estimate generalisation accuracy in training large-scale data. Hence, the hold-out procedure is a better choice from the perspective of training costs.

The response time of KMSVM (or SVM) classifiers is affected by the number of support vectors according to the representation of KMSVM (SVM) classifiers in equation (2). Hence, model selection is implemented according to the tradeoff between testing accuracy and response time (the number of support vectors).

There are some strategies for searching the input parameters, among which the simplest one is grid search. The time to find good input parameters by grid search is not much more than by advanced methods when there are only two input parameters in SVM. When the grid-search method is adopted, trying exponentially growing sequences of C and γ is a practical method to identify good input parameters, e.g., $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$ (Hsu et al., 2003). For example, model selection is implemented on the Adult-7 database and the number of support vectors and testing accuracy from the different combinations of input parameters are recorded in Tables 2 and 3. In Table 2, we cannot find a combination of the kernel parameter γ and the penalty factor C, which reduces support vectors to fewer than 5000. This implies that only in the input parameter space constructed by the kernel parameter γ and the penalty factor C, it is difficult to search a combination of input parameters to get many fewer support vectors so that the response time can be improved.

The KMSVM algorithm extends the 2-dimensional input parameter space in SVM to a 3-dimensional input parameter space constructed by the kernel parameter γ, the penalty factor C, and the compression rate CR. Grid search is time consuming when it is run for more than two input parameters. In order to avoid high computational cost of grid search, this paper proposes a heuristic searching strategy to determine the good combination of parameter values in this 3-dimensioanl input parameter space.

- Several good combinations of input parameters are determined by grid search in SVM, and then for these different combinations of input parameters the different numbers of clusters is tested to find a final combination of the three input parameters trading-off testing accuracy and response time. For example, in the example about the Adult-7 database, firstly two combinations of C and γ are identified in SVM, e.g., $C_1 = 2^{15}$ and $\gamma_I = 2^{-11}$ (testing accuracy is 84.9% and the number of support vectors is 5670), $C_2 = 2^9$ and $\gamma_2 = 2^{-11}$ (the testing accuracy is 84.9% and the number of support vectors is 5781).
- For these two combinations of C and γ, the compression rate CR is searched among some values, e.g., 10, 20, …, 60. Finally the combination of $C_2 = 2^9$, $\gamma_2 = 2^{-11}$, and $CR = 60$ is determined because of the good testing accuracy and a less response time (see Tables 2–4).

Table 2 Number of support vectors for different combinations of the kernel parameter γ and the penalty factor C. The different value (–5, –3, ..., 15) of $\log_2 C$ is in every row and the different value (–15, –13, ..., 3) of $\log_2 \gamma$ is in every column. '–'s in the table indicate that these input parameters are ignored because of too long training time

	–15	*–13*	*–11*	*–9*	*–7*	*–5*	*–3*	*–1*	*1*	*3*
–5	7,836	7,836	7,836	7,840	7,842	7,427	7,109	10,744	14,640	14,639
–3	7,836	7,836	7,842	7,846	7,240	6,497	6,399	10,738	14,641	14,640
–1	7,836	7,844	7,848	7,201	6,417	6,042	6,186	10,562	14,634	14,634
1	7,841	7,849	7,186	6,403	5,998	5,863	6,353	10,610	14,333	14,339
3	7,850	7,188	6,403	6,001	5,837	5,842	6,678	10,497	14,345	–
5	7,185	6,404	5,999	5,845	5,756	5,943	6,565	10,513	14,341	–
7	6,460	5,995	5,840	5,774	5,709	6,085	6,341	10,499	14,338	–
9	5,996	5,840	5,781	5,719	5,729	6,058	6,279	10,497	14,338	–
11	5,833	5,783	5,753	5,681	5,831	5,808	6,269	10,493	–	–
13	5,777	5,760	5,715	5,653	5,876	5,528	6,251	10,516	–	–
15	5,760	5,751	5,670	5,643	5,792	5,441	6,252	10,504	–	–

This heuristic strategy is feasible for the fact that the optimal combination of the kernel parameter and the penalty factor determined in SVM can be approximately regarded as the optimal one in the KMSVM algorithm because the K-means clustering technique can almost preserve the underlying structure and distribution of the original data. This is also verified by the experiments below, i.e. there is an insignificant degradation of testing accuracy when the same combination of the kernel parameter γ and the penalty factor C are applied with a different number of clusters (e.g., $CR = 10$, 20) in the KMSVM algorithm.

Table 3 Testing accuracy for different combinations of the kernel parameter γ and the penalty factor C. The different value (–5, –3, …, 15) of $\log_2 C$ is in every row and the different value (–15, –13, …, 3) of $\log_2 \gamma$ is in every column. '–'s in the table mean that these input parameters are ignored because of too long training time

	–15	*–13*	*–11*	*–9*	*–7*	*–5*	*–3*	*–1*	*1*	*3*
–5	76.2	76.2	76.2	76.2	77.2	83.0	82.5	76.9	76.2	76.2
–3	76.2	76.2	76.2	77.6	83.6	83.7	83.6	80.7	76.2	76.2
–1	76.2	76.2	77.7	83.7	84.2	84.3	84.6	83.0	77.0	77.0
1	76.2	77.8	83.8	83.3	84.7	84.7	84.6	82.7	77.7	77.7
3	77.8	83.8	84.4	84.8	84.7	84.9	83.1	82.1	77.7	–
5	83.8	84.4	84.9	84.9	84.8	84.4	81.1	82.1	77.7	–
7	84.4	84.8	84.8	84.8	84.8	83.0	80.4	82.1	77.7	–
9	84.8	84.8	84.9	84.9	84.5	81.1	80.4	82.1	77.7	–
11	84.8	84.9	84.8	84.9	83.6	79.5	80.4	82.1	–	–
13	84.9	84.9	84.9	84.6	82.2	79.0	80.4	82.1	–	–
15	84.8	84.8	84.9	83.7	80.2	78.8	80.4	82.1	–	–

Table 4 Adjusting the number of clusters on the Adult database. $CR = 1$ means that the classifier is built by SVM and $CR = 10 \ldots 60$ means that the classifiers are built by the KMSVM algorithm with different numbers of clusters

Methods		*SVM*	*KMSVM*					
Compression rate		*1*	*10*	*20*	*30*	*40*	*50*	*60*
Adult–4	Number of SVs	1747	193	*108*	71	61	48	*41*
	Response time (s)	45.3	7.6	*5.2*	4.3	3.6	3.2	*3.0*
	Testing accuracy (%)	84.5	82.2	*82.1*	81.1	81.4	81.7	*82.8*
Adult–5	Number of SVs	2311	250	141	97	*73*	61	52
	Response time (s)	56.6	8.9	5.9	4.6	*3.9*	3.5	3.2
	Testing accuracy (%)	84.5	83.6	83.3	83.3	*82.4*	82.1	81.6
Adult–6	Number of SVs	4023	437	235	156	124	*99*	85
	Response time (s)	79.7	11.7	7.3	5.5	4.6	*3.9*	3.6
	Testing accuracy (%)	84.7	84.0	82.9	80.3	81.0	*82.2*	80.9
Adult–7	Number of SVs	5781	602	319	212	170	137	*112*
	Response time (s)	88.0	12.3	7.4	5.4	4.6	4.0	*3.5*
	Testing accuracy (%)	84.9	84.2	83.9	83.1	82.7	82.0	*82.1*

In unsupervised learning, the number of clusters is subjectively determined by users according to their domain knowledge. There are some objective criteria and strategies to determine the number of clusters (e.g., the hold-out criterion and the heuristic strategy proposed in this paper) when the K-means clustering technique is combined with SVM to solve the problems in the supervised learning. The experiments in Section 4 show that the KMSVM algorithm can find a good combination of input parameters to greatly reduce the number of support vectors and response time of classifiers and maintain a similar testing accuracy to SVM.

4 Experiments

This section presents experiments on two real-world data sets to verity the effectiveness of the KMSVM algorithm. The hold-out procedure and the heuristic searching strategy in this paper are adopted to determine three input parameters. Testing accuracy and response time are used to measure the performance of classifiers. The Gaussian RBF in equation (3) is used as the kernel function in our experiments. The experiments are performed on a Pentium4 CPU with 256 MB memory. SVM and the KMSVM algorithm are implemented on the software ‘*LIBSVM 2.4*’ from www.csie.ntu.edu.tw/~cjlin/libsvm. The results show that it is possible for the KMSVM algorithm to build classifiers with a high response speed and a similar testing accuracy compared with SVM.

Experiment 1

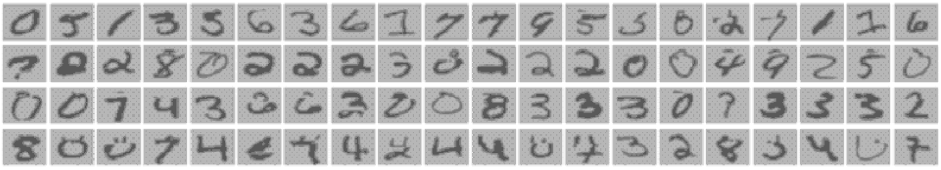
This experiment is run on the Adult data from the UCI machine learning database repository. The database is separated into training data and testing data, and there are nine groups (Adult-1a, Adult-1b, …, Adult-9a, Adult-9b) of training and testing data in the database. The goal of this data mining task is to predict whether a household has an income greater than US$ 50,000 (pl. query) using the census form of the household. This experiment selects four groups of data (Adult-4a, Adult-4b … Adult-7a, Adult-7b) to train and test SVM classifiers and the KMSVM classifier. Performance of these classifiers is evaluated by testing accuracy and response time. For simplicity, the values of input parameters determined on the Adult-7 data are also applied to other data sets, i.e., the Gaussian RBF kernel parameter $\gamma = 2^{-11}$ and the penalty factor $C = 2^{9}$ in this experiment.

The results recorded in Table 4 show a bigger compression rate (that is, fewer cluster centres), fewer support vectors, and higher response speed. For example, on the Adult-7 data, when the support vectors are reduced more than 50 times (112/5861), there is only a 2–3% discrepancy of testing accuracy between the KMSVM classifiers and the SVM classifiers (82.1% vs. 84.9%) and the response time of the KMSVM classifiers is about 25 times less than that of the SVM classifiers (3.5s vs. 88s). Furthermore, we have discovered that for the four Adult databases, fewer than 100 support vectors can make testing accuracy up to 82%. This implies that about 100 support vectors may be enough for this data mining task.

Experiment 2

This experiment is performed on the US Postal Service (USPS) database (LeCun et al., 1990), which is often used to evaluate the performance of classifiers based on SVM. The USPS database includes a lot of real hand-written digits from ‘1’ to ‘10’ (see Figure 4). The USPS database has been separated into a training set of 7291 samples and a testing set of 2007 samples. In this experiment, we use the Gaussian RBF kernel parameter $\gamma = 0.0078125$ (2^{-7}) and the penalty factor $C = 5$. The results in Table 5 show that compared with SVM, the KMSVM algorithm can build classifiers with about 3.7 times less response time to separate ten hand-written digits with a similar testing accuracy (about a 2–3% discrepancy of testing accuracy). Hence, this experiment can also verify the effectiveness of the KMSVM algorithm.

Figure 4 Normal and atypical hand-written digits



Source: Wang et al. (2003).

Table 5 Adjusting the number of clusters on the USPS database

Methods	*SVM*			*KMSVM*			
Compression rate	1	10	20	30	40	50	60
Number of SVs	*1450*	324	*190*	137	126	99	83
Response time (s)	*17.0*	5.9	*4.6*	4.1	3.9	3.6	3.6
Testing accuracy (%)	*95.5*	93.4	*92.53*	91.88	91.98	91.38	90.88

5 Conclusion

SVM has been applied to solve some BI problems, but the response time of SVM classifiers still needs to improve when applied into real-time BI systems. This paper has proposed a new algorithm, called KMSVM, to build classifiers by combining the K-means clustering technique with SVM. Besides, this paper has given a criterion and strategy to determine the input parameters in the KMSVM algorithm. Experiments on the real-world databases have shown that compared with SVM, the KMSVM algorithm can build classifiers with both a higher response speed and a similar testing accuracy. This could be useful for real-time BI systems, such as stock market surveillance and network intrusion detection. In the future, the KMSVM algorithm should be verified on more real-time BI databases.

References

Boser, B.E., Guyon, I.M. and Vapnik, V. (1992) 'A training algorithm for optimal margin classifiers', *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, ACM Press.

Duda, R.O. and Hart, P.E. (1972) *Pattern Classification and Scene Analysis*, Wiley, New York.

Furey, T.S., Cristianini, N., Bednarski, D.W., Schummer, M. and Haussler, D. (2000) 'Support vector machine classification and validation of cancer tissue samples using microarray expression data', *Bioinformatics*, Vol. 16, No. 10, pp.906–914.

Ganapathiraju, A., Hamaker, J. and Picone, J. (2004) 'Applications of support vector machines to speech recognition', *IEEE Transactions on Signal Processing*, Vol. 52, No. 8, pp.2348–2355.

Hsu, C.W., Chang, C.C. and Lin, C.J. (2003) *A Practical Guide to Support Vector Classification*, Working Paper.

Joachims, T. (1998) 'Text categorization with support vector machines: learning with many relevant features', *Proceedings of the 10th European Conference on Machine Learning*.

Langford, J. (2000) 'PAC bounds for hold-out procedures', *Cross-Validation, Bootstrap and Model Selection Workshop*, *Advances in Neural Information Processing Systems*.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. and Jackel, L.J. (1990) 'Handwritten digit recognition with back-propagation network', in Touretzky, D.S (Ed.): *Advances in Neural Information Processing Systems*, Vol. 2, pp.396–404.

Müller, K.R., Mika, S., Rätsch, G., Tsuda, K. and Schölkopf, B. (2001) 'An introduction to kernel-based learning algorithms', *IEEE Transactions on Neural Networks*, Vol. 12, No. 2, pp.181–201.

Osuna, E., Freund, R. and Girosi, F. (1997) 'Training support vector machines: an application to face detection', *Proceedings of the 1997 IEEE Conference on Computer Vision and Pattern Recognition*.

Vapnik, V. (1998) *Statistical Learning Theory*, Wiley, New York.

Viaene, S., Baesens, B., Van Gestel, T., Suykens, J.A.K., Van den Poel, D., Vanthienen, J., De Moor, B. and Dedene, G. (2001) 'Knowledge discovery in a direct marketing case using least squares support vector machines', *International Journal of Intelligent Systems*, Vol. 16, No. 9, pp.1023–1036.

Wang, J.Q., Tao, Q. and Wang, J. (2002) 'Kernel projection algorithm for large-scale SVM problems', *Journal of Computer Science and Technology*, Vol. 17, No. 5, pp.556–564.

Wang, J.Q., Zhang, C.Q., Wu, X.D., Qi, H.W. and Wang, J. (2003) 'SVM-OD: a new SVM algorithm for outlier detection', *Foundations and New Directions of Data Mining Workshop in IEEE International Conference of Data Mining*.